Protecting children from harms online

Volume 5: What should services do to mitigate the risks of online harms to children?

Consultation

Published 08 May 2024
Closing date for responses: 17 July 2024

Contents

Section

13. Our proposals for the Children's Safety Codes

Volume 5 outlines draft measures we propose providers of services likely to be accessed by children could take to comply with their child safety duties in the Online Safety Act ('the Act'). These are set out in the draft Children's Safety Codes in Annexes 7 and 8. These measures will be finalised following consultation.

Services likely to be accessed by children are required by the Act to use proportionate safety measures to keep them safe. Our draft Children's Safety Codes provide a set of safety measures that online services can take to help them meet their duties under the Act. Services can decide to comply with their duties by taking different measures to those in the Codes. However, they will need to be able to demonstrate that they offer the appropriate level of safety for children.

Our draft Codes bring together a broad package of safety measures that aim to protect children online. They also work alongside the other pillars of the Online Safety regime to collectively improve safety online for everyone, especially children.

These measures are based on our assessment of the risks that children face online from content designated as harmful to children in the Act – see Volume 3. Evidence shows content harmful to children is highly prevalent online across all types of services and many UK children encounter it. The impact of harmful content and activity can be wide-ranging and severe. Across content types, children's emotional wellbeing is affected and at worst, harmful content and activity can contribute to loss of life. Much of what we know about the risk of harm to children comes from engaging with children. As part of our research, children told us what they want and need to ensure they can live a safer life online, including the measures they would like to see service providers implement. We have also drawn together substantial input from the online sector, as well as children's organisations, academics, independent researchers, and other public bodies. These insights helped inform our analysis of possible protections.

Our proposals

There is no single fix-all measure that service providers can take to protect children online. Safety measures need to work together to help create an overall safer experience for children. We are proposing a set of safety measures in our draft Children's Safety Codes, that will work together to achieve safer experiences for children online.

We are proposing more than 40 safety measures in our draft Children's Safety Codes, in these broad areas:

- **Robust age checks.** We expect much greater use of age assurance, so services know which of their users are children. All services which do not ban harmful content, and those at higher risk of it being shared on their service, should implement highly effective age-checks to prevent children from seeing it.
- **Safer algorithms**. Recommender systems – algorithms which provide personalised recommendations to users - are children's main pathway to harm online. Under our proposals, any service which operates a recommender system and is at higher risk of harmful content should identify who their child users are and configure their algorithms to filter out the most harmful content from children's feeds and reduce the visibility of other harmful content.

- **Effective moderation**. All user-to-user services should have content moderation systems and processes that ensure swift action is taken against content harmful to children. Search services should also have appropriate moderation systems and, where large search services believe a user to be a child, a 'safe search' setting which children should not be able to turn off should filter out the most harmful content.
- **Strong governance and accountability**. Proposed measures here include having a named person as accountable for compliance with the children's safety duties; an annual senior-body review of all risk management activities relating to children's safety; and an employee Code of Conduct that sets standards for employees around protecting children.
- **More choice and support for children**. This includes ensuring clear and accessible information for children and carers, with easy-to-use reporting and complaints processes, and giving children tools and support to help them stay safe.

We expect these measures to make a big difference to children's online experiences. For example:

- Children will not normally be able to access pornography.
- Children will be protected from seeing, and being recommended, potentially harmful content.
- Children will not be added to group chats without their consent.
- It will be easier for children to complain when they see harmful content, and they can be more confident that their complaints will be acted on.

Over time, as we gather more information on how the risks of harm to children online evolve and how our proposals are impacting this, we expect that we will consider whether it is appropriate to add further measures to future iterations of the Children's Safety Codes. We have identified, and set out later in this section, a number of areas where we believe additional measures could play a potential significant role in delivering protections for children online or where technology is evolving and our understanding of the risk of harm to children is emerging. We also explain how we will be seeking input from children via a programme of deliberative engagement. In addition, we welcome expressions of interest, in particular from service providers, to work with Ofcom's Behavioural Insights hub to better understand 'what works' through testing and trialling the design of potential future measures with children.[1]

Consultation Questions

Proposed measures

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

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 costs of such measures, including any results from trialling or testing of measures.

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.

[1] Please express interest via email: Behavioural.insights@ofcom.org.uk

Introduction to Volume 5

13.1 Volume 5 sets out the proposed steps service providers should take to keep children safe online.

13.2 The Online Safety Act 2023 ('the Act') requires all Part 3 service providers to carry out a children's access assessment to establish whether their services are likely to be accessed by children. Our proposals for the draft Children's Access Assessment are discussed in Volume 2 of this consultation. Providers of services likely to be accessed by children must carry out a children's risk assessment. Our proposals for the Children's Risk Assessment Guidance are discussed in Volume 4 of this consultation.

13.3 Providers of services likely to be accessed by children must take appropriate and proportionate measures to effectively mitigate the risks that their services pose to children, as identified in their children's risk assessments. This volume sets out our proposed measures for service providers to mitigate risks to children and meet the children's safety and reporting and complaints duties in the Act.[2] When finalised these will form the Children's Safety Codes**,** drafts of which are published separately as Annex 7 for user-to-user ('U2U') services and Annex 8 for search services. Proposed measures apply to different services based on their level of risk, and in some cases based also on size and functionalities.

13.4 Services that choose to implement the measures we recommend in Ofcom's Children's Safety Codes will be treated as complying with the relevant children's safety as well as their reporting and complaints duties.[3] This means that Ofcom will not take enforcement action against them for breach of that duty if those measures have been implemented. This is sometimes described as a "safe harbour." However, the Act does not require that service providers adopt the measures set out in the Children's Safety Codes, and service providers may choose to comply with their duties in an alternative way that is proportionate to their circumstances.[4]

13.5 The rest of this section briefly describes our package of proposed measures for this first iteration of the Children's Safety Codes, how this will protect children online and the wider impacts of our proposals on users and services. To conclude, we set out areas for possible further work, where we also seek feedback on the proposed priorities.

13.6 In Section 13 in this volume, we explain the approach we have taken to develop our proposed measures, and our framework for assessing their impact on children and adults, and on services. In Sections 14-21, we consider in more detail each of the groups of measures we are proposing for inclusion in the Children's Safety Codes. Proposed governance and accountability measures are included in Volume 4 at Section 11.

[2] Sections 12, 29, 20, 31, 21, 32 of the Act.
[3] Section 49(1) of the Act
[4] If service providers choose to comply with their children's safety and reporting and complaints duties in another way, the Act provides that, they must have regard to the importance of protecting users' right to freedom of expression within the law, and to the importance of protecting users from breaches of relevant privacy laws: see section 49(5). Where providers do take alternative measures, they must keep a record of what they have done and explain how they think the relevant safety duties have been met. This is described in more detail in Annex 6: Guidance on record keeping and review in our 2023 Illegal Harms Consultation.

The risks to children from harmful content online

13.7 The measures we propose to include in the Children's Safety Codes are informed by our assessment of the risks presented by content harmful to children. This is the focus of our analysis of the causes and impacts of harm to children as presented in Volume 3 of this consultation. We have prioritised addressing the most significant risks identified in our analysis and those required by the Act.

13.8 We focus on three categories of content, within which sit several kinds of harmful content, as specified in the Act. We set these out in Definition box 1 below. For readability, we refer to individual kinds of harmful content using shorthand in bold:

Definition box 1: Summary of types of harmful content, as defined in the Act

Primary Priority Content ('PPC')	Pornographic content **Suicide and self-harm content:** Content which encourages, promotes or provides instructions for suicide or encourages, promotes or provides instructions for an act of deliberate self-injury. **Eating disorder content:** Content which encourages, promotes or provides instructions for an eating disorder or behaviours associated with an eating disorder.
Priority Content ('PC')	**Abuse and hate content:** Content which is abusive and which targets any of the following characteristics— (a) race, (b) religion, (c) sex, (d) sexual orientation, (e) disability, or (f) gender reassignment. Content which incites hatred against people— (a) of a particular race, religion, sex or sexual orientation, (b) who have a disability, or (c) who have the characteristic of gender reassignment. Bullying content.[5] **Violent content:** Content which encourages, promotes or provides instructions for an act of serious violence against a person. Content which— (a) depicts real or realistic serious violence against a person; (b) depicts the real or realistic serious injury of a person in graphic detail. Content which— (a) depicts real or realistic serious violence against an animal; (b) depicts the real or realistic serious injury of an animal in graphic detail; (c) realistically depicts serious violence against a fictional creature or the serious injury of a fictional creature in graphic detail. **Harmful substances content:** Content which encourages a person to ingest, inject, inhale or in any other way self-administer— (a) a physically harmful substance; (b) a substance in such a quantity as to be physically harmful. **Dangerous stunts and challenges content:** Content which encourages, promotes or provides instructions for a challenge or stunt highly likely to result in serious injury to the person who does it or to someone else.

[5] Many research sources use the term 'cyberbullying' within their analysis when referring to bullying content and behaviour online. In line with the Act, we use the term 'bullying content' or 'bullying online' throughout this section.

Definition box 1: Summary of types of harmful content, as defined in the Act	
Non designated content ('NDC')	Content, which is not primary priority content or priority content, of a kind which presents a material risk of significant harm to an appreciable number of children in the United Kingdom.

13.9 Our risk assessment shows that content harmful to children is present online across all types of services and many UK children encounter it.[6] In a four-week period, 62% of children aged 13-17 report encounter PPC/PC online.[7] Research also found children consider violent content 'unavoidable' online.[8] Pornographic content is particularly pervasive in the online lives of children with nearly two-thirds of children and young adults (13-19) reporting ever having seen pornographic content. Of these, most (73%) had done so by the age of 15, around a quarter by age 11 and one in ten as young as 9.[9] Some children encounter several types of harmful content – especially those spending the most time online.[10]

13.10 The impacts of viewing harmful content are wide-ranging and can be severe. Across content types, children's emotional wellbeing is affected.[11] It can lead to feelings of anxiety, shame or guilt; can discourage children from expressing themselves online; or even risk children adopting attitudes and behaviours that cause harm to peers and communities.[12] At worst, harmful content can contribute to loss of life.[13] While all children are at risk, harmful content can also disproportionately affect certain groups. For instance, the number of girls aged 13-21 who have been subject to abusive or hateful comments online – specifically 'sexist comments' – has almost tripled in ten years from 20% in 2013 to 57% in 2023.[14]

---

[6] See Volume 3.
[7] Ofcom, 2023. Online Experiences Tracker. Note: Fieldwork was conducted in June-July 2023, so 'in the last four weeks' refers to responses in this time period.
[8] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.
[9] Children's Commissioner (2023) 'A lot of it is actually just abuse' Young people and pornography. [accessed 14 June 2023]
[10] Internet Matters found that over a fifth of the children who spent the most time online (the top quartile) reported experiencing five or more potential harms online. The Index is based on responses to a detailed survey of 1,000 children aged 9-15 and their parents, conducted during summer 2022. Source: Internet Matters, 2023. Children's Wellbeing in a Digital World: Year Two Index Report 2023. [accessed 24 April 2024]
[11] For example, regardless of their own experience, children report feelings of anxiety, shame, guilt and fear on encountering content promoting eating disorders. Ofcom, 2024. Online Content: Qualitative Research. Experiences of children encountering online content relating to eating disorders, self-harm and suicide.
[12] For example, evidence links violent content to specific behaviours related to violence, such as leading children to perceive it as normal to carry knives – see Volume 3.
[13] The coroner's report for 14-year-old Molly Russell concluded that watching high volumes of content promoting suicide and self-harm had contributed to her death by suicide. (The Coroner concluded that it was likely that the material viewed by Molly, who was already suffering with a depressive illness, and vulnerable due to her age, affected her mental health in a negative way and contributed to her death in a more than minimal way. Source: Courts and Tribunals Judiciary, 2022. Molly Russell: Prevention of future deaths report - Courts and Tribunals Judiciary, 13 October 2022. [accessed 16 April 2024].) The inquest into the death by suicide of 14-year-old Mia Janin found that she had been experiencing bullying online. (BBC, 2024. Mia Janin took own life after bullying – inquest. [accessed 14 February 2024].) There are also several examples from around the world of children losing their lives after attempting challenges circulating online. (See Impacts section in 'Content promoting dangerous stunts and challenges' - Volume 3).
[14] Girlguiding, 2023. Girls Attitudes Survey 2023: Girls' lives over 15 years. [accessed 24 April 2024]

13.11 Our analysis of the causes and impacts of harm has highlighted that some types of services – including social media and video-sharing services – play a particularly prominent role in disseminating harmful content. Some service characteristics are also particularly important in the dissemination of harmful content. This includes recommender systems, which, in their current form, expose children to many categories of harmful content and often in high volumes and combinations of harmful content. Further analysis of other risk factors (such as features and functionalities of a service, its user base or business model) are explored in the draft Children's Register of Risk at Section 7. Where we identify functionalities as posing risks to children this is not to say that these functionalities are in and of themselves harmful or that they don't serve a useful function, but that they can be deployed in a problematic manner.

13.12 Much of what we know about the risk of harm to children comes from engaging with children. As part of our research, children also told us what they want and need to ensure they can live a safer life online, including the measures they would like to see platforms implement. These insights helped inform our analysis of possible protections.

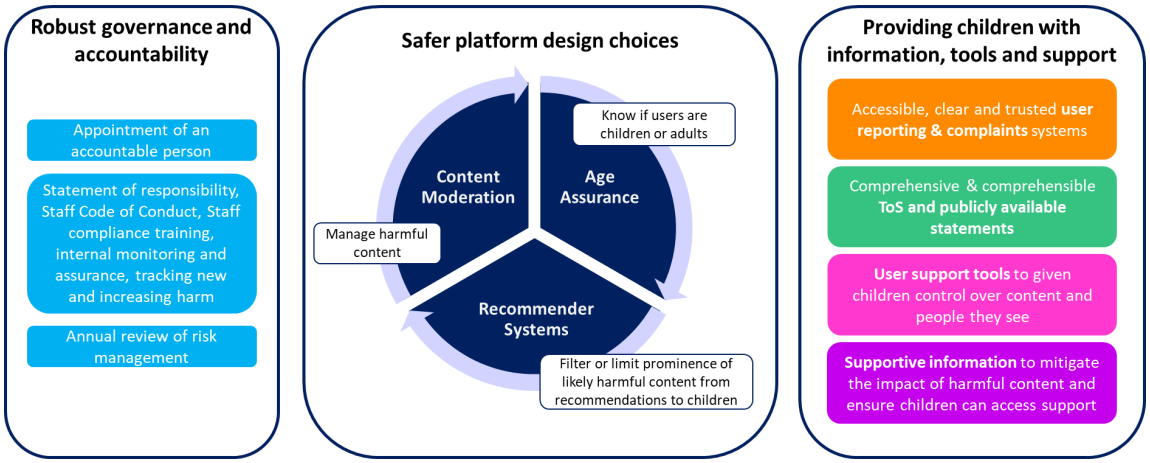
13.13 Existing protections for children, where available, are fragmented. Different services offer children very different experiences which do not necessarily correlate to the level of risk that they face on those services.

Our proposals to protect children online

13.14 The first Children's Safety Codes will set the baseline of protections that should be in place across the industry to protect children. They will form a strong set of **foundations to protect children online.**

13.15 To determine which of the measures from the Children's Safety Codes to apply, providers of services likely to be accessed by children will need to identify what specific risks of harm they pose to children using the Children's Risk Assessment Guidance (Section 11). Ofcom's Guidance on Content Harmful to Children (Section 8) describes the kinds of content that Ofcom considers falling within the relevant definitions of PPC and PC under the Act, along with a non-exhaustive list of examples of content that may and may not fall within those categories. The Children's Register of Risk (Section 7) sets out how we find those harms manifesting in the current environment. This, alongside the findings of their children's risk assessment, will largely inform what measures will be appropriate for service providers to implement to mitigate risks to children and meet the duties in the Act.

Figure 13.1. Package of proposed measures for Children's Safety Codes



13.16 There is no single fix-all measure that services can take to protect children online. Safety measures need to work together to help create an overall safer experience for children. We have proposed a set of safety measures within our draft Children's Safety Codes that will work together to achieve safer experiences for children online. These cover three broad areas, which we discuss in turn below:

- **Robust governance and accountability –** ensuring service providers have appropriate senior oversight and accountability for children's safety online;
- **Safer platform design choices –** making sure services understand their users' ages and keep children safe, including ensuring recommender systems and content moderation operate effectively to prevent harm to children;
- **Providing children with information, tools and support** – ensuring service providers provide clear and accessible information to children and carers, making sure reporting and complaints functions are easy-to-use, and giving children tools and support to help them stay safe.

13.17 Our impact assessments for each proposed measure, as well as in combination, are set out at Sections 15-23. This includes consideration of the risk of harm to children that could be addressed by our proposed measures, the effectiveness of the measures in mitigating said risk, the costs of implementing measures as well as possible impacts on the rights and user experience of children and adults. We set out which services we propose the measures should apply to. In many cases, measures that pose significant costs would still apply to providers of smaller services, if they meet the relevant criteria for each measure. We anticipate this means some smaller services may stop allowing PPC/PC on their services or stop serving the UK altogether. We think the measures are nonetheless proportionate to ensure children are protected from harm on the services where they are at most risk.

Robust governance and accountability

13.18 Strong governance and accountability are crucial to service providers' efforts in protecting children online. By governance and accountability, we mean the structures and processes organisations use to ensure there is adequate oversight of decision-making, roles and responsibilities, and effective reporting and review mechanisms.

13.19 We are, therefore, proposing measures for how service providers should approach governance and accountability in relation to protecting children online. We propose that all U2U, and search services name a person accountable for compliance with the child safety duties. For services that are either large or multi-risk (or both), we also propose a package of more comprehensive measures to ensure internal monitoring and assurance functions are in place over how risks to children are managed and evaluated, as well as written statements of responsibility, a Code of Conduct for employees and training for relevant staff. We discuss these in Section 11 in Volume 4.

13.20 These complement the related guidance for providers in our draft Risk Assessment Guidance, discussed in Volume 4, Section 12 and published as Annex 6. We think that the totality of these measures will ensure there is a high level of senior oversight of how service providers are handling and mitigating risks of harm to children – and help make sure services are designed and operated in ways that effectively mitigate those risks.

13.21 Our approach is consistent with our Illegal Harms Consultation. This means service providers who must comply with both illegal content safety duties and children's safety duties can choose to adopt a single process that covers both areas.

Safer platform design choices

13.22 We are also proposing a range of safety measures that focus on service providers ensuring they make foundational design choices, so children have safer online experiences. These cover three broad topics:

- understanding which users are children so that those children can be kept safe;
- ensuring recommender systems do not operate to harm children; and
- making sure content moderation systems operate effectively.

Understanding which users are children so they can be protected online

13.23 We do not want children to be denied their rights or enjoying the benefits of being online, but they should be protected from exposure to harmful content.

13.24 We are proposing broader use of age assurance so that services know which of their users are children, so they have a safe experience online – see Section 15. We are proposing 'age assurance' to be used by services that pose risks to children. Where we recommend services use age assurance, we propose that they use what we refer to as 'highly effective age assurance'.

13.25 This is age assurance that is highly effective at correctly determining whether or not a user is a child. We propose that the age assurance used should fulfil the criteria of technical accuracy, robustness, reliability, and fairness. We have published our draft guidance on highly effective age assurance at Annex 10.

13.26 Our proposals recognise that age assurance is not a silver bullet and will not be the only effective solution to protect children in all scenarios. We are therefore recommending that highly effective age assurance be used in the areas where it can have the most impact in protecting children online. We have also been mindful of the need to preserve the rights of adult users in accessing legal content. Ultimately, our proposals are designed to protect children from encountering harmful content, and to strengthen the effectiveness of other measures we set out in Volume 5 (which might rely on knowing the age of a user).

Ensuring recommender systems do not operate to harm children

13.27 Recommender systems are a primary method for sharing users' content across services. Recommender systems use algorithms to curate and determine how content is shown to users (including children) based on their characteristics, inferred interests, and behaviour. They are generally designed to make the service more appealing to users, by showing them content that the recommender system determines is likely to be of interest to them.

13.28 Evidence shows that recommender systems are a key pathway for children to encounter harmful content, including suicide, self-harm and eating disorder content, violent content, and pornographic content. They also play a part in narrowing down the type of content presented to a user, which can lead to increasingly harmful content recommendations as well as exposing users to cumulative harm over time through repeated exposure to harmful content or harmful combinations of content.

13.29 We are therefore proposing three safety measures targeting the design and operation of recommender systems to ensure children are protected from encountering harmful content on recommended feeds and have more control over the content that is recommended to them – see Section 20. Our proposals recommend that U2U services operating a recommender system, and posing a risk of exposing children to content harmful to them, follow a precautionary approach to content shown in children's feed. This is achieved through excluding content likely to be PPC (Measure RS1) and limiting the prominence of content likely to be PC (Measure RS2). On large risky services, children should also be offered more control, allowing them to indicate if they do not want to continue to see certain types of content (Measure RS3). This corresponds with views from some children in our research, primarily in relation to PPC, who expressed a desire for more control over the content they see and the ability to directly impact their personal feed to avoid content they want to see less of.

13.30 We think these propose measures will work together to mitigate the risks of harm that recommender systems pose to children, in particular the risk of exposure to cumulative harm.

13.31 Our proposed measures on recommender systems are different under this consultation to the Illegal Harms Consultation. In our 2023 consultation, we proposed a safety measure for service providers to collect metrics on recommender systems and assess whether any changes are likely to increase user exposure to illegal content. This proposal from the Illegal Harms Consultation will also help protect children from illegal content.

Making sure moderation systems work effectively

13.32 Content moderation is the process by which a service reviews content to decide how it should be treated on its service. If it is content harmful to children and access to it should therefore be restricted, services should take steps to ensure children are prevented or protected from encountering it. Content moderation can be done automatically using technology, by human moderators, or a combination of the two. Content moderation plays a hugely important role in keeping users safe from harm - especially children.

13.33 Evidence shows that content harmful to children is available on many services at scale, and that children are regularly exposed to it.[15] This suggests that services' current efforts to

[15] See sub-section 'The risks to children from harmful content online' above.

protect children from harmful content (including content moderation) are not working well enough.

13.34 We ultimately expect all user-to-user and search services to put in place effective systems to address content that is harmful to children and take swift action to protect them from it. This might include ensuring such content is not shown to children or taking the content down if it is not permitted. For services that are either large or multi-risk to children (or both), we propose a package of more comprehensive measures to ensure that these processes are fit for purpose given the more complex risk environment these services operate in. This set of proposals do not include expectations on the use of automated tools to detect and review content. However, we are aware that large services often do so to handle the scale of content and are exploring how to incorporate measures on automated tools into our Codes.

13.35 Our proposed measures target the effectiveness of content moderation systems on user-to-user and search services – see Sections 16 and 17. They are focused on making sure services have in place effective systems and processes to act on content that is harmful to children, clear policies on what is allowed, adequate moderation resources, and effective systems to prioritise how content is moderated. We think these measures will support more effective content moderation systems and processes, in turn reducing the likelihood that children encounter harmful content.

Providing children with information, tools and support

13.36 We are also proposing a range of safety measures that focus on service providers providing children with information, tools, and support that will help to keep them safer online. These cover three broad topics:

- having clear terms of service and publicly available statements;
- making sure children can easily report content and make complaints; and
- providing children with tools and support to help them stay safe.

Having clear terms of service and publicly available statements

13.37 Terms of service (terms) and publicly available statements (statements) typically lay out the rights and responsibilities that a service provider and the users of their service have towards one another. Terms and statements tend to contain information about how a service functions, including who is allowed to use the service, rules for using the service and how users will be protected from harm on the service.

13.38 Children and the adults who care for them need to refer to terms or statements if they want to understand the provisions providers have in place to help protect them. If this information is not provided by a service or if the information is presented in a confusing or inaccessible way, children and carers might not be able to make informed choices about whether to use a service. In addition, it might be difficult for them to know what content is allowed and recognise content that is harmful and report it. This could contribute to the prolonged presence of content harmful to children on a service.

13.39 We are therefore proposing that all user-to-user and search services should ensure their terms of service and/or publicly available statements are comprehensive, clear, and accessible for children and the adults that care for them – see Section 19. Children should be able to understand what content is allowed on a service and what is not – and this should be presented as clearly as possible.

13.40 We think our proposals will increase children's knowledge and confidence in using online services, including any means the service provides for them to control their own user experience. This in turn should help children to recognise and submit a report or complaint if they are exposed to harmful content online. This should contribute to a safer online environment for children.

13.41 Our proposals are broadly consistent with the measures proposed in our Illegal Harms Consultation. However, we are proposing a new measure for providers of Category 1 and 2A services – that they should summarise the findings of their most recent children's risk assessment in their terms or statement.

13.42 We are also proposing an equivalent measure (Measure 6AA) for Category 1 and 2A services relating to their illegal content risk assessment to add to the proposals set out in our Illegal Harms Consultation. This measure recommends that Category 1 and 2A services should summarise the findings of their most recent illegal content risk assessment in their terms or statement.

Making sure children can easily report content and make complaints

13.43 User reporting and complaints allow users – including children – to make service providers aware of when harmful content is present on their service, or when content has been mistakenly removed or restricted. They both play an important role in protecting children online and protecting users' rights.

13.44 While many services already have reporting and complaints functions available to users, our evidence suggests that children do not think these are always accessible, easy to use and transparent. This can discourage people from using these functions, including children.

13.45 In our Illegal Harms Consultation, we proposed a range of measures to help providers meet their duties under the Act, in relation to the design and operation of their complaints and reporting processes. In this consultation, we are proposing measures to build on those already proposed in the Illegal Harms Consultation, to help services meet their duties. These measures, in summary, require services likely to be accessed by children to both have easy to access, easy to use and transparent complaints processes, and acknowledge and take appropriate action in response to complaints – see Section 18.

13.46 Our proposed measures refer to 'complaints', which include user reports, appeals and other types of complaints, such as complaints about a service not complying with its duties to protect children. User reports are a specific type of complaint about content, submitted through a reporting tool. Appeals are complaints by users who believe a service has made an incorrect decision about a piece of content.

13.47 We know that many providers operate a single complaints process for various types of complaints. We have taken this into account when assessing what measures to propose in the Children's Safety Codes. Many of the proposed measures align with measures in the draft Illegal Content Codes. However, Measures UR2 and UR3 include additional elements, which we provisionally think should be included in both the Children's Safety Codes and the Illegal Content Codes. These are recommendations that services should explain to complainants when they make a complaint what, if any, information they will provide, and services should include information about the resolution of complaints in the acknowledgement they send to complainants.

13.48 The expansion of these measures has been proposed because of new evidence relating to children's concerns about the confidentiality of services' complaints processes, and how the

lack of a satisfactory communication from services about a complaint could reduce trust in the complaints process overall. We therefore propose that services should do more to explain to users how their complaints procedures operate, for the purposes of developing more transparent complaint mechanisms.

13.49 We believe these measures will ensure services have effective complaints procedures in place, which will help them take steps to protect children from encountering harmful content and improve any systems they use to detect harmful content. This will ensure services can be made safer for children, accountable and respectful of user rights.

Providing children with tools and support to stay safe

13.50 Many services have functionalities that allow users to connect with one another, such as group messaging or comment sections. These functionalities can pose risks to children, as they can allow users to expose children to harmful content or activity without their consent.

13.51 We have proposed user support measures which we believe will give children more control over their online experience and help them stay safe online – see Sections 21 and 22. For user-to-user services, we are proposing:

- **user support tools** that will enable children to have more control over their interactions on services that pose a risk of harm, by giving them the option to decline group invites, block and mute user accounts, or disable comments on their own posts; and
- **user support materials** for children to both assist their understanding of how they can restrict certain types of online interactions that may put them at risk of harm and to support them when they report, post, or search for certain types of harmful content. These measures apply depending on a service's risk level and size.

13.52 These measures broadly mirror those that we proposed relating to user support in our Illegal Harms Consultation. Measure US4 is also an adapted version of a measure in our Illegal Harms Consultation - we are proposing for certain types of services to provide information to child users when they restrict interactions with other accounts or content.[16] We will consider whether to apply this additional element to our Illegal Harms Codes ahead of finalising them. We are also proposing three user support measures for the Children's Safety Codes that do not have an equivalent in our proposed Illegal Harms Codes.[17]

13.53 For search services, we are also proposing a measure to provide crisis prevention information in response to search requests for known primary priority content (which includes self-harm and suicide content). Crisis prevention information includes help and support such as helplines and supportive information from reputable organisations. This measure is also consistent with what we proposed in our Illegal Harms Consultation.

13.54 We believe these measures will give children more control over their online interactions and provide added support while online to help keep them safe.

[16] Under the corresponding measure in our Illegal Harms Consultation (Measure 7B), service providers would provide information to children when they are taking action against another user, but not when they are taking action against content.

[17] Measures US1, US5 and US6

The role of children's voices in our codes proposals

13.55 Together with this consultation, we have launched a deliberative engagement programme and are developing child-friendly materials to use in workshops with over 100 children aged 8+ to include their views in our proposals. This builds on our previous work with around 80 children to gather their views on the Illegal Harms measures designed to protect children from online grooming. This direct engagement is an important part of ensuring that we are proposing robust measures that have a meaningful impact on their user experience and the harms they encounter.

13.56 This engagement builds on our ongoing **research programme**[18] to develop our knowledge on the risks of harm to children, what works to keep them safe, and their views on what measures can protect them online. [19] In April 2024, we published Ofcom's Online Safety Research Agenda which highlights children's online experiences and the impact of harms as areas of particular interest to us. One of our key priorities as part of our research programme is direct engagement with children, building on the experience from our long-standing media literacy programme of research and more recently the significant evidence base on the risks and harms children are facing online.

13.57 The planned components of our children's research and engagement programme include:

- a Children's Online Research Panel which will facilitate ongoing engagement with children in a variety of ways;
- a Children's Online Safety Tracker to monitor risks and harms, and attitudes and experiences of online safety measures;
- our Online Passive Measurement tool to better understand the online platforms and services being used by children;
- and further behavioural trials among under 18s to better understand how to positively influence children's decision making online.

13.58 This research will be complemented with direct work with regulated service providers to help us understand, assess, and drive improvements and understand what measures might work to protect children from harm.

13.59 We invite stakeholder expressions of interest to collaborate with Ofcom's Behavioural Insights Hub on testing potential future measures with children, to help develop our evidence base.[20]

Areas for future Children's Safety Codes measures

13.60 The proposals in this consultation mark a vital first step toward safeguarding children online. We're committed to continuously refining our strategies based on a dynamic understanding of both the digital landscape and children's experiences on the internet. Through an active programme of research and ongoing dialogues with services —including targeted information requests—we aim to keep our approach fresh and effective.

[18] For more information about our research programme please visit Protection of children online, research - Ofcom

[19] Across our research programme, children shared their experiences and told us what they want and need to be protected online.

[20] Please express interest via email: Behavioural.insights@ofcom.org.uk

13.61 We've pinpointed several critical areas that demand urgent attention and possibly further action. These include using automated content moderation to detect illegal and harmful content on a large scale, addressing the risks children face from emerging generative AI technologies, and tackling features that entice children to increase their screen time. Furthermore, we're exploring more tailored protection strategies for different age groups and examining how parental controls can not only empower parents but also enhance their children's safety online.

Automated content moderation

13.62 The identification of content harmful to children at scale is key to protecting children online and is integral to the effectiveness of protections of users from Illegal Harms, such as the detection of child sexual abuse material. For many larger services, the use of proactive technology (notably automated content classifiers) plays a key role in identifying illegal and harmful content at scale.

13.63 The Act requires us to have regard to the degree of accuracy, effectiveness, and lack of bias achieved by any technologies that we propose to recommend, which would enable services to comply with their safety duties.[21] The Act also requires that we must be satisfied that the technology in question is proportionate to the risk of harm the measure is designed to safeguard against.[22] These principles reflect the risk from proactive technology of a disproportionate interference with users' fundamental rights to privacy, freedom of expression and access to information.

13.64 For this consultation, we considered recommending specific automated technologies, such as keyword detection for the identification of content harmful to children and nudity detection technology. Our view was that these technologies by themselves might not currently be sufficiently sophisticated to accurately detect harmful content to children and could result in the suppression of relevant sources of information/content that are not harmful to children.[23] While we decided to recommend the use of hash matching technology for some harms based on the evidence available on their accuracy, effectiveness and lack of bias, we did not propose other types of automated content moderation technology for the detection of wider illegal harms.

13.65 Despite potential limitations of specific technologies, we are planning an additional consultation later this year on how automated detection tools can be used to mitigate the risk of illegal harms and content harmful to children. This will include previously undetected child sexual abuse material and content encouraging suicide and self-harm. These proposals will draw on our growing technical evidence base and will complement the existing measures set out in our draft Codes of Practice.

[21] Paragraph 13(6) of Schedule 4 to the Act
[22] Paragraph 13(5) of Schedule 4 to the Act
[23] While keyword detection may be used as part of a wider content moderation systems and processes, we considered additional layers of technology would be needed to effectively detect PPC.

Generative Artificial Intelligence

13.66 Generative artificial intelligence ('GenAI') is a development in artificial intelligence that refers to machine learning models that can create new content in response to a user's prompt. These models can be used to produce text, images, audio, videos and code. Our research has found that children are using and interacting with GenAI on both U2U and search services, as well as on standalone GenAI applications, including chatbots that can provide information and recommendations, to image generators that allow them to create avatars, stickers, and immersive video gaming content.[24]

13.67 There is emerging evidence that GenAI can facilitate the creation of content harmful to children across several types of harmful content defined in the Act, including pornography, content promoting eating disorders and bullying (Section 7.14 in Volume 3). The Online Safety Act is "tech neutral". This means that GenAI-created content which is harmful to children and is shared on a U2U service or is presented in search service results needs to be treated in the same way as other forms of content harmful to children under our proposed measures, and in accordance with the Act.

13.68 As services deploy new technologies, including features and functionalities powered by GenAI, they must consider the risk that they pose to children and how to ensure they have the appropriate mitigations in place to address those risks and will need to update their risk assessment after any major change to their service, including the introduction of changes to GenAI powered functionalities.

13.69 We are undertaking a programme of research to explore the effectiveness of measures to identify and address harmful AI-generated content on online platforms, including red teaming and deepfake detection.[25] We plan to publish our findings in June 2024.

Impact of choice architecture

13.70 There is a substantive body of research that explores how service design and functionalities are applied to influence online behaviours, including the presentation and placement of choices and the design of interfaces. [26] These design features are sometimes referred to as "persuasive design."

13.71 Some online choice architecture practices can be designed to encourage users (including children) into maximising the frequency or time spent on a service (e.g. 'infinite scrolling'; auto-play features; affirmation-based functionalities; alerts and notifications). On the other hand, choice architecture can also help nudge users (including children) into safer behaviours (e.g. reporting flags on front page; clear and accessible links to support materials).

13.72 As part of our analysis in the draft Children's Register of Risk, we considered the impact of service design and functionalities that affect the amount of time that children spend online, and engagement with services (Section 7.13 in Volume 3). Our evidence shows that the risk to children of encountering harmful content on a service increases with the time and

[24] Ofcom, 2023. Online Nation.

[25] **Red teaming:** a mode of content evaluation targeted at finding vulnerabilities in AI systems and applications; **Deepfake detection**: techniques that can be used to identify deceptive or misleading content that has been manipulated or created outright using AI or related digital techniques.

[26] Competition & Markets Authority, April 2022. Evidence review of Online Choice Architecture and consumer and competition harm

frequency of use. Our proposed measures in this code are designed to reduce the risk of children encountering harmful content, including from functionalities like the recommender system that can help amplify exposure to harmful content. As part of our scoping exercise, we considered the role of functionalities such as autoplay in amplifying the risk of harm but decided not to propose any specific recommendations at this stage given the more limited evidence on the role of autoplay in amplifying exposure of children to harmful content compared to other functionalities like recommender systems.

13.73 As part of this consultation, we look to establish if there are residual concerns with choice architecture in scope of the Act that need to be considered for consideration in the Children's Safety Codes.

Children of different ages

13.74 In this first iteration of our Children's Safety Codes we are focusing on proposals that will result in safer, more protected experiences for all children, which are defined in the Act as users under the age of 18. The Act also requires all children to be prevented from encountering PPC and expects children "in age groups judged to be at risk of harm" to be protected from other harmful content.

13.75 We recognise that age is a key factor that will affect children's expectations and experiences of being online and our research indicates that certain online behaviours vary by age and developmental stage. However, there is currently limited evidence on the specific impact of harms to children in different age groups and limited existing technologies that can reliably identify children of different ages. Given these limitations, our proposals focus at this stage on setting the expectation of protections for all children under the age of 18. Services are required to take into consideration children in different age groups as part of their risk assessment and we have provided guidance on how best to meet this requirement. We encourage services to tailor their experiences to children in different age groups, based on their understanding of their user base and the risks that their services pose.

13.76 As part of this consultation, we want to understand if our proposals are likely to have a disproportionate impact on children in different age groups, especially in relation to PC and in relation to older children and their rights and freedom to access information, and how we might be able to build more flexibility to mitigate these negative impacts while ensuring they receive the right protections from harmful content.[27]

[27] For more information about our research programme please visit Protection of children online, research - Ofcom.

Parental controls

13.77 Parental tools can support parents and carers to exercise a degree of choice over the online experiences of their children and can have a beneficial impact on the online experiences of those children. The evidence about the effectiveness and uptake of existing parental tools as a way of increasing children's safety online is currently limited.[28] Existing research, based on our engagement with stakeholders, has also shown that services have very different approaches to parental controls and offer different functionalities, which are often limited.

13.78 Our current set of proposals focus on ensuring services meet their obligations to build experiences for children that are safer by design, in line with the duties in the Act which place responsibility for protecting child users explicitly on service providers. Based on responses to this consultation and our own additional research we will continue to explore the complementary role that parental controls can play as part of future iterations of the Children's Safety Codes in supporting safer experiences online of children in different age groups.

[28] For example, see Ofcom, 2023. How video-sharing platforms (VSPs) protect children from encountering harmful videos.

14. Developing the Children's Safety Codes: Our framework

This section explains the approach we have taken to develop our draft Children's Safety Codes.

The Children's Safety Codes explain how providers of services likely to be accessed by children can comply with the children's safety and reporting and complaints duties in the Online Safety Act. The Children's Safety Codes are one of several sets of Codes of Practice that Ofcom is developing, through public consultation, to keep users safe in line with duties on service providers under the Act. We consulted on our proposed measures for services to comply with illegal content duties in our 2023 Illegal Harms Consultation.

The draft Children's Safety Codes and the draft Illegal Content Codes (published in November 2023) are consistent with each other. In some areas, our proposals in the Children's Safety Codes build on, and in some cases closely mirror, measures proposed in the Illegal Content Codes. In other areas, we are proposing additional measures which are intended to specifically mitigate the risks to children from content that is harmful to them but is not illegal. Service providers in scope of both sets of Codes should consider recommended measures together once finalised.

In line with the approach to the draft Illegal Content Codes, and as required by the Act, the measures we recommend must be proportionate. We consider the impact of each of our proposed measures individually and in combination with other proposed measures, including, where relevant, measures we proposed for the Illegal Content Codes. This is rooted in evidence and includes consideration of:

- The risk of harm to children that could be addressed by our proposed measures, including **scale and severity;**
- The **effectiveness** of the proposed measures in **mitigating risks of harm;**
- The **costs of implementing measures**, both direct and indirect; and
- Possible **impacts on the rights and user experience of children and adults.**

Consultation questions

25. Do you agree with our approach to developing the proposed measures for the Children's Safety Codes? Please explain why.
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.
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?
28. Do you agree with our definition of 'large' and with how we apply this in our recommendations?
29. Do you agree with our definition of 'multi-risk' and with how we apply this in our recommendations?
30. Do you agree with the proposed measures that we recommend for all services, even those that are smaller and low-risk?

Purpose of the Children's Safety Codes

14.1 The Online Safety Act 2023 ('the Act') requires Ofcom to prepare and issue Codes of Practice ('Codes') for user-to-user ('U2U') and search services. When finalised, these codes will set out measures we recommend service providers implement to comply with the duties in the Act.

Definition box 1: What are U2U and search services?	
U2U services	An internet service on which users of the service can generate, upload and/or share content, which can then be encountered by other users of the service.
Search services	An internet service that is, or includes, a search engine. Includes **general search services** which enable users to search any contents of the web and return results. Services might do this by relying on their own indexing using bots to find content across the web, building an index of URLs and using algorithms to rank content. Also includes services that are **vertical search services**, which present users with results only from selected websites with which they have a contract, an API or similar technology.

14.2 In this volume, we set out for consultation proposed measures for inclusion in Codes of Practice for services to comply with the children's safety and reporting and complaints duties ('children's safety duties').[29] In essence these duties require services to use measures to protect children from content harmful to children, namely Primary Priority Content (PPC), Priority Content (PC) and Non-designated Content (NDC).[30] For the detail of the duties that apply to U2U and search services please refer to Volume 1, Section 2, and Annex 13.

14.3 Only services likely to be accessed by children are subject to the children's safety duties in the Act. As discussed in Volume 2, all Part 3 services are required to carry out children's access assessments to determine whether they are likely to be accessed by children. The duties, and therefore the Codes of Practice, relate to the design and operation of services in the UK or as it affects UK users of the service. The duties apply to providers of such services, even if they are based outside the UK. Services that operate across different jurisdictions have a choice: they may choose to apply all the safety protections that the Act requires for all their users, no matter where in the world; or target them specifically to users in the UK.

14.4 Once finalised, our proposed measures will become the Children's Safety Codes. Drafts of the Children's Safety Codes are published at Annex 7 (measures recommended for U2U services) and Annex 8 (measures recommended for search services).

14.5 Services likely to be accessed by children, and which choose to implement the measures we recommend in the Children's Safety Codes, will be considered as complying with relevant duties. This means that Ofcom will not take enforcement action against services for breach of a duty if the relevant measures have been implemented.

14.6 Services may choose to comply with the children's safety duties in a different way from the measures we recommend in our Codes. In doing so, services must have regard to the

[29] The children's safety duties are set out in sections 12 and 29, and the reporting and complaints duties in sections 20, 31, 21 and 32 of the Act.
[30] See Definition box 1 in Section 13 of Volume 5.

importance of protecting users' rights to freedom of expression within the law and relevant privacy laws – see also Volume 1, Section 2. Services must also keep a written record of any measures they take to comply with the relevant duties and explain how they think any alternative measures taken have met their safety duties.[31] More detail on this is included in our draft record keeping guidance, published with our 2023 Illegal Harms Consultation in Annex 6.

Scope of the Children's Safety Codes

14.7 Ofcom is developing the Children's Safety Codes in parallel with other Codes for compliance with different duties in the Act. The Act requires Ofcom to prepare and issue three sets of Codes for Part 3 services (U2U and search services), namely a Code covering the illegal content safety duties, including terrorism content, a Code on CSEA content, and one or more Codes of Practice for the purposes of compliance with other relevant duties. We consider that physically separate Code documents for each of these would be repetitive and potentially confusing for stakeholders. Instead, we seek to present the three sets of Codes in two groups (each containing two documents – one covering U2U and another covering search services) organised as follows:

- **Codes of Practice relating to Illegal Content duties** – including CSEA, terrorism content and other priority illegal content, content reporting duties for illegal content and complaints procedures (sections 10, 27, 20, 31, 21 and 32).
- **Codes of Practice relating to Children's Safety duties** – including the safety duties protecting children, content reporting duties for content harmful to children and complaints procedures (sections 12, 20, 21, 29, 31, and 32).

14.8 We are also developing additional proposals for duties that providers of categorised services will need to comply with. This will form the third phase of implementing the Act, building on the first phase which relates to illegal harms.

14.9 We think organising the Codes in this way will help provide clarity. In particular, given the children's safety duties distinctly only apply to services likely to be accessed by children, as opposed to the illegal content duties which apply to all services, we consider it appropriate to have bespoke Codes of Practice for the protection of children.

14.10 While the draft Illegal Content Codes and the draft Children's Safety Codes are distinct, they will still be closely intertwined.

14.11 Both the Illegal Content Codes and the Children's Safety Codes protect children. The illegal content safety duties protect children from illegal content and the children's safety duties protect children from harmful content other than illegal content. Accordingly, several measures proposed for the Children's Safety Codes build on proposals in the Illegal Content Codes. In the areas of user reporting and complaints, governance and accountability, content moderation (U2U and Search), user support and terms of service, some of our proposed measures closely mirror proposals for the Illegal Content Codes.

14.12 Where relevant, we have mirrored the proposed Illegal Content measure and tailored the content so that services can meet their children's safety duties. Even where the outcome of a proposed measure for the Illegal Content Codes may be the same as a proposal for the

[31] The record keeping duties are set out at section 23 of the Act.

Children's Safety Codes, we consider it important to mirror the measure for the Children's Safety Codes to give clarity to service providers likely to be accessed by children as to the full scope of measures we propose for the safe harbour with the children's safety duties.

14.13 Where appropriate, we allow flexibility for services to leverage common systems and processes to implement both sets of measures. For example, a service may choose to operate a single reporting system that caters for user complaints related to illegal content and content harmful to children.

14.14 We are consulting on some additional proposals for the Illegal Content Codes alongside the Children's Safety Codes (see Annex 9). In some areas, new child-specific evidence has given us a more granular understanding of possible measures to protect children from harmful content. In some cases, this new evidence also supports the case for additional protections for children from illegal content, or has resulted in revisions to our existing proposals, in the Illegal Content Codes. The Act is clear in its objective for services to provide a higher standard of protection for children than adults.[32]

14.15 Specifically, we are proposing additional measures for the Illegal Content Codes concerning:

- Ease of access, use and transparency of complaints systems (Section 18)
- The acknowledgement of complaints (Section 18)
- The substance of information to be included in Terms of Service or Publicly Available Statement (Section 19)
- Materials for volunteer moderators (Section 16).

Coordinating across Ofcom consultations and statements

14.16 We expect to publish our statements on the Illegal Content Codes around the end of 2024 and on the Children's Safety Codes in early 2025. We are developing the Illegal Harms and Children's Safety Codes in parallel. This means that we are receiving stakeholder feedback and progressing our work on the different Codes at the same time. Some of the responses we have received to our Illegal Harms Consultation may also be relevant for our proposals on the Children's Safety Codes, and vice versa. We will take into account relevant responses to our Illegal Harms Consultation, as well as all responses to this consultation, as we prepare our statement on the Children's Safety Codes.

14.17 In relation to a few points, we have already been able to take into account relevant response to our Illegal Harms Consultation. These are exceptions and we have not made any judgments on the merits of other responses to the Illegal Harms Consultation.

14.18 If you have already responded to the Illegal Harms Consultation and would like us to consider some or all your response in relation to this consultation, please let us know.

Our approach to developing recommended measures

14.19 All the measures proposed in the Codes have been developed in line with Ofcom's duties set out in legislation - in particular the Communications Act 2003 ('CA 2003'), the Online Safety Act 2023 ('the Act'), and the Public Sector Equality Duty. We discuss these duties in detail in Annex 13, Annex 14, and Section 24.

[32] Schedule 4 paragraphs 4(a)(vi); 5(a)(v) of the Act.

14.20 The CA 2003 sets out a number of duties Ofcom must fulfil in exercising our regulatory functions, including our principal duty to further the interests of citizens in relation to communication matters and further the interests of consumers in relevant markets where appropriate by promoting competition.[33]

14.21 In carrying out our functions, we are required to secure the adequate protection of citizens from harm presented by content on regulated services. In developing our Children's Safety Code proposals, we have had regard to factors including, but not limited, to the following:

- The risk of harm to citizens presented by content on regulated services.
- The need for a higher level of protection for children than for adults.
- The need to be clear to providers how they may comply with their duties.
- The need to exercise our functions so as to secure that providers may comply with the duties using measures which are proportionate to the size or capacity of the provider and the level of risk of harm presented by the service.
- The desirability of promoting the use by providers of technologies which are designed to reduce the risk of harm to citizens presented by content on services.[34]
- As appropriate, the desirability of promoting competition and encouraging investment and innovation in relevant markets.[35]

14.22 Further, Schedule 4 of the Act sets out online safety objectives. Accordingly, we must ensure that measures described in Codes are compatible with these objectives. The full detail is set out in Section 24. It includes taking account of the needs of different kinds of users and the overall user base, effectiveness, and proportionality. Schedule 4 also requires us to include measures in the Codes in each of the categories of measures contained within services' safety duties in sections 12(8) and 29(4).

14.23 In line with our additional duties under section 3(4) of the CA 2003, we have also considered the vulnerability of children and of others whose circumstances put them in need of special protection; the needs of persons with disabilities, the elderly and of those on low incomes; the opinions of consumers and of members of the public generally; and the different interests of persons in the different parts of the United Kingdom and of the different ethnic communities within the United Kingdom.[36]

Our evidence base

14.24 Under the Act,[37] we are required to carry out impact assessments when preparing a Code of Practice (or amendment to a Code of Practice), which also includes an assessment of the impact on small and micro businesses. To do this, and in line with requirements, principles and objectives, our proposals must be evidence-based. However, we are developing measures for a sector without previous direct regulation. This means that the volume of evidence and independent analysis can be limited in some areas.

14.25 Over the past two years, we have sought to fill evidence gaps to help us understand what measures might be proportionate and effective in protecting children across the operation and design of services.

[33] Section 3(1) CA 2003
[34] Section 3(4A) CA 2003
[35] Section 3(4)(b) and (d) CA 2003
[36] Section 3(4)(h) - (l) CA 2003
[37] Section 93(4) of the Act

14.26 We conducted a large programme of research – both ourselves and by commissioning independent research.[38] Much of what we learned about the risk of harm to children comes from engaging with children. As part of our research, children told us what they want and need to ensure they can live a safe life online, including the measures they would like to see service providers implement.

14.27 We also conducted extensive stakeholder engagement and sought third-party input to build our evidence base. We held a call for evidence on risks of harms to children online and how they can be mitigated ('2023 Protection of Children Call for Evidence') and received evidence from a wide range of stakeholders including civil society organisations and service providers. We followed this up with seven roundtable discussions hosted across the UK including Belfast, Edinburgh and London. These were attended by a range of stakeholders including organisations focused on specific harms to children, as well as one dedicated session with industry stakeholders. We have not yet formally requested information from service providers as our information gathering powers only came into effect in January 2024.

14.28 This evidence has helped us build an understanding and identify areas of focus for the development of measures for the Children's Safety Codes. Our ambition is to provide clarity to services for how to deliver on the children's safety duties as quickly as possible so that we can drive improvements in children's online lives sooner.

14.29 This is only the first iteration of the Children's Safety Codes. Over time and as we work to fill evidence gaps, we intend to iterate and add to the measures we are proposing in this consultation. In Section 13, we discuss our immediate priority areas for further evidence gathering.

The impact assessment framework

14.30 We consider a wide range of impacts when we assess which measures to recommend, how to design these measures and the kinds of services in scope of each measure. At the heart of our assessment is the extent to which our package of measures can reduce the risks that children face when using regulated services. This allows us to identify which measures are most effective at protecting children and to target those measures towards services where children face the greatest risks.

14.31 At the same time, the Act requires that any adverse effects of our measures are appropriate and proportionate to our objective of improving children's safety. Our proposed package of measures is designed to achieve this objective, without undermining the important benefits that online services in scope of the Act deliver to UK citizens. The potential adverse effects will vary depending on the measure and can include:

a) impacts on user rights, e.g. privacy or freedom of expression;
b) impacts on user experience, e.g. adding friction to the user journey; and
c) costs for regulated services, which may indirectly affect users, for instance if service quality degrades due to higher costs, or if competition, innovation and choice are reduced.

14.32 We have sought to quantify impacts where feasible, but there are limits to the extent to which we have been able to do so for a range of reasons, including:

[38] Ofcom's online research is published here: https://www.ofcom.org.uk/research-and-data/online-research.

- Some impacts are difficult to quantify due to a lack of robust evidence, including about services' current systems, costs and effectiveness of existing measures.
- Some impacts are of a less tangible nature and more challenging to quantify fully in economic terms, such as non-economic impacts of children's exposure to harmful content on wellbeing or even loss of life. While not necessarily quantified, such impacts can be very material and have had a strong influence on our decisions.
- The broad scope of the regime means there is uncertainty around the number of services in scope, the prevalence of relevant characteristics across services, and the resources available to different kinds of services. These factors can influence how different services may be impacted by our proposed measures.

14.33 As a result, there is typically a degree of uncertainty over the magnitude of impact of any individual measure and the package of proposed measures. In particular, it has not been possible to quantify the benefits of measures in terms of harm reduction. Nonetheless, a qualitative assessment of the effectiveness of measures in reducing the risks to children online is at the core of our proposals, given the very broad and severe harms associated with content harmful to children. This includes the range of psychological, emotional and other harms to the children affected and to those around them, as well as the economic costs to society. Moreover, even where we do include estimates of costs, these should be interpreted as indicative. Given the uncertainties highlighted above, it is possible that the cost for a given service may be below or above the ranges provided, depending on its specific context.

14.34 Working with imperfect evidence means that we face uncertainty when making our recommendations, with some decisions being finely balanced. Online services in scope of the Act, and the technologies they use, are evolving rapidly – and new harms may emerge. There is a need for prompt action to protect children online and a clear risk that children will not be protected if we only recommend measures where we have extensive and definitive direct evidence of effectiveness. Therefore, some of our proposed measures are based on an assessment of more limited or indirect evidence of impact, and reliance on logic-based rationales. We exercise regulatory judgement in prioritising measures which, on balance, we consider can materially improve children's safety online. In some cases, where we provisionally conclude that certain measures should not be recommended at this stage, or only recommended for some services but not others, we intend to consider this further as we review the responses to this consultation and as part of our future work.

14.35 There are also certain measures where we deliberately do not assess impacts in detail. This is the case for measures that closely reflect specific requirements in the Act which all services in scope of the children's safety duties must follow. Examples of these include duties to have a complaints process, or to include certain information in terms of service and publicly available statements in a clear and accessible way. In such cases, our proposed measures represent the minimum necessary for services to comply with those particular duties and we allow discretion to services in terms of *how* they implement those measures. For such measures, we consider that impacts result from the Act itself rather than any exercise of our regulatory discretion, so we consider the measures to be proportionate without requiring a detailed examination of their impacts.

The effectiveness of the measure in addressing risk of harm to children

14.36 In each of the sections, we have set out the risk of harm to children that we are seeking to address through our proposed measures. To inform this, we draw on our evidence of the risks of harm to children as set out in our analysis of the causes and impacts of harm to children in Volume 3. We have carefully considered how the proposed measures will reduce the risk of harm to children we have identified, including any evidence of services currently practicing the measure or version of it, and their technical feasibility. Where we highlight current practices, this does not represent an endorsement of the service's approach, nor does it mean that the service is meeting the requirements of the code or an indication of compliance.

14.37 Our evidence indicates that some of the cross-cutting measures that we propose – for example, the governance and accountability and content moderation measures – have the potential to reduce harm in relation to all kinds of content harmful to children, by ensuring that services operating in a more complex risk environment employ suitably sophisticated systems and processes. Other measures target risks associated with specific kinds of content harmful to children, or with specific functionalities that have been shown to pose risks to children, such as recommender systems. These complement the cross-cutting measures, by addressing specific risk factors associated with end-user functionalities.

14.38 We use this evidence to assess qualitatively how, and to what extent, different measures contribute to safer experiences for children online. This helps us to design and prioritise measures in line with our objectives.

Rights assessment

14.39 In accordance with our obligations under the Human Rights Act 1998, Ofcom must consider the impacts that our regulatory proposals could have on human rights set out in the European Convention on Human Rights (ECHR) and ensure that they are compatible with these rights. For each proposed measure, we consider human rights implications, in particular the right to freedom of expression (Article 10 ECHR), the right to freedom of association (Article 11) and the right to privacy (Article 8 ECHR). We have sought to secure that any such interference with adults' and children's relevant rights, is proportionate to the legitimate objective of the Act of protecting children from content harmful to them. We also recognise that our proposed measures could help to protect individuals from harms of various kinds (including in particular the duties aimed at protecting children from harm which are the key focus of this consultation, as well as the duties which apply to illegal content and activity) which reflect the decision of the UK Parliament that UK users, and UK children in particular, should be proportionately protected from all the harms concerned. We discuss our overarching approach for how we do this in further detail in Volume 1, Section 2.

14.40 Our approach is consistent with the principles of the United Nations Convention on the Rights of the Child, and in particular the provision that the best interests of the child should be a primary consideration in all regulatory actions concerning children. This is reflected in the children's safety duties and the way that the Act requires Ofcom to seek to secure a higher level of protection for children than for adults.

14.41 Along with the right to privacy conferred by Article 8 ECHR, there are domestic laws that are relevant to this right. Services will need to ensure they comply with data protection law which includes the Data Protection Act 2018, the UK General Data Protection Regulations (UK GDPR) and where relevant, the Privacy and Electronic Communications (EC Directive)

Regulations (PECR). Users' rights to data protection are regulated by the Information Commissioner's Office (ICO). The ICO has a range of data protection compliance guidance[39] which we encourage services to consult. In particular, services should familiarise themselves with the ICO's Children's Code, the ICO Commissioner's Opinion on Age Assurance and their guidance on Online Safety and data protection.[40]

Further impacts

14.42 Where relevant, we also consider potential further impacts on children and adults. For example, this could include added frictions to user journeys and access to services or content, as well as any other costs or possible unintended consequences.

Equality impact assessment and Welsh language

14.43 In line with our public sector equality duties, we have considered the equality impacts of our proposed measures and draft Guidance to comply with our duties under the Equality Act 2010 and the Northern Ireland Act 1998 and set out our understanding of any particular impacts on protected groups in the UK.[41] We consider that some of our proposals would have a positive impact on certain groups. In addition to impacts in relation to our draft Codes of Practice proposals, Ofcom's proposed guidance on content harmful to children is also likely to have positive equality impacts on certain groups. In formulating our proposals in this consultation, where relevant and to the extent we have discretion to do so in the exercise of our functions, we have considered the potential impacts on opportunities to use the Welsh language and treating the Welsh language no less favourably than English, in accordance with the Welsh language standards. Our considerations in relation to equality impacts and Welsh language are set out at Annex 14.

Impacts on services

14.44 Ofcom is under a duty to carry out an impact assessment in carrying out functions where proposals are important.[42] A proposal to prepare a Code of Practice is deemed important under the CA 2003.[43] Before implementing such a proposal, Ofcom must carry out and publish an impact assessment.[44] The CA 2003 requires us to consider the impact of our measures on services of different sizes and capacity including an assessment of the likely impact of implementing the proposal on small businesses and micro businesses.[45]

14.45 Impacts on services are an important consideration to ensure that more costly requirements are justified, even where they could negatively affect users. For example, if a high-cost burden on services reduces investment in areas other than user safety or (in the most extreme cases) drives some services to stop operating in the UK, this means that both children and adults can no longer benefit from such services or new innovations. This does

[39] For further guidance, see please see the ICO for organisations.
[40] ICO, Age Appropriate Design Code (which we refer to as the 'Children's Code'), 2022, Commissioner's Opinion on Age Assurance for the Children's Code and Online safety and data protection.
[41] We have given careful consideration as to whether the proposals will have a particular impact on persons sharing protected characteristics (including race, age, disability, sex, sexual orientation, gender reassignment, pregnancy and maternity, marriage and civil partnership and religion or belief in the UK and also dependents, and political opinion in Northern Ireland), and in particular whether they may discriminate against such persons or impact on equality of opportunity or good relations. Impact assessments at Annex 14.
[42] Section 7(1) of the CA 2003.
[43] Section 7(2A) of the CA 2003 (as amended by the Act).
[44] Section 7(3)(a) of the CA 2003.
[45] Section 7(4A) of the CA 2003 (as amended by the Act).

not mean that services should not fulfil their duties to keep children safe because it is costly. Considering the cost impact on services aims to meet the child safety requirements under the Act without unduly undermining investment in high-quality online services that UK users can enjoy, including children.

14.46 Our assessment of impacts on services considers:

a) Direct costs of implementing the measures, including any one-off costs and any ongoing costs. Where these costs are quantified, these often rely on salary and other assumptions as detailed in Annex 12: Further detail on economic assumptions and analysis.
b) Indirect costs or risks, where applicable, such as any possibility of reduced user engagement and revenue.

14.47 We consider costs on a per-service basis, which allows us to assess implications of our measures for services of different sizes and capacity,[46] including small and micro businesses. We employ commonly-used definitions across many government bodies, where a small business is defined as one with 10-49 full-time employees and a micro business is one with fewer than 10 full-time employees (in either case, this may include employees not based in the UK).[47]

14.48 In many instances, we allow some flexibility in how services can practically implement our recommendations, to ensure services can take an approach that is appropriate and proportionate to their circumstances. In those areas where we are proposing to be more prescriptive around the details of the practical implementation, this is because we consider it necessary for the measures to have the intended effect, and our assessment and discussion of cost is typically more detailed in such cases.

14.49 Our analysis focuses on what the costs would be for those services that are not currently undertaking the measures. Some services may already have the same or similar measures we are proposing in place, including where services are in scope of similar measures proposed in our Illegal Harms Consultation. In our analysis we identify potential cost synergies in such cases, which can reduce the incremental cost of implementing our proposed measures to protect children.

Which providers we propose each measure should apply to

14.50 We recognise that the size, capacity, functionalities, user base and risks of online services in scope of the children's safety duties differ widely. For this reason, we have not taken a one-size-fits-all approach and a key part of our decision-making concerns which kinds of services each measure should apply to. The measures we are proposing may have different impacts on services of different kinds and sizes. As a result, some of our recommendations apply to all Part 3 services (including where this is required by the Act itself); others may apply to different kinds of services, depending on whether they meet one or more criteria based on:

[46] See Annex 12: Further detail on economic assumptions and analysis.
[47] We appreciate that not all Government bodies use exactly the same definitions. For example, some also refer to revenue and assets. The definition we propose is consistent with that used by the Regulatory Policy Committee. It would not make a material difference to our impact assessment if another common definition of small and micro business (such as that consistent with the Companies Act 2006) were used instead. Source: Regulatory Policy Committee, 2019. Small and Micro Business Assessments: guidance for departments, with case history examples, August 2019.

a) Whether the service is a U2U or search service, with further distinctions made between different kinds of search service (general search or vertical search).
b) The outcome of the service's latest risk assessment, with respect to the level of risk for each kind of content harmful to children and the number of risks identified.
c) The size of the service, in terms of its UK user base.
d) The functionalities or other relevant characteristics of a service (e.g. use of recommender systems or community moderation).

14.51 Our framework for defining the kinds of services in scope of each measure, including with reference to size and risk thresholds, is broadly similar to that adopted for our Illegal Harms Consultation. We have not yet processed all responses to our 2023 Illegal Harms Consultation and it is possible that in light of these responses we may make adjustments to this framework in future.

14.52 There are measures that apply to services even if they are low-risk, meaning that they do not have a medium or high risk for any kind of content harmful to children. These reflect steps that we expect all services should take to comply with the children's safety duties, including with respect to their terms of service, user reporting processes and content moderation processes. However, services that do pose significant risks to children are expected to take additional steps, as we summarise below.

14.53 Overall, we consider the nature and level of risk that a service poses to children to be the main driver for whether to recommend measures for that service. The benefits to children from a measure being implemented will generally be greater where a service poses higher risk of harm to children. Focusing most measures on these services is consistent with ensuring proportionality.

14.54 Various measures are recommended for services that have medium or high risk for at least one kind of content harmful to children, from a defined subset of kinds of content relevant to each measure. These measures are intended to target specific risk factors, often linked to end-user functionalities (such as recommender systems or group chats), strengthening the protection of children from specific harms, on the services where such harms may arise. For example, some user support measures aim to give children more control over their online experiences, through tools that allow them to block user accounts and disable comments, which can reduce risk of harm related to cyberbullying, abuse and hate content.

14.55 Further proposed measures are recommended for services that are multi-risk for content harmful to children, meaning that they have medium or high risk for two or more kinds of content harmful to children (i.e. at least two across the four kinds of PPC, eight kinds of PC and any kinds of NDC where applicable). These measures primarily focus on governance and content moderation, aiming to ensure that multi-risk services adopt more sophisticated systems and processes, enabling them to manage multiple risks effectively given their more complex risk environment. These more general measures contribute to reducing harm related to *any* kind of content harmful to children, complementing the measures discussed in the previous paragraph, which target *specific* harms.

14.56 We also propose a small set of measures for large services only, where there is significant scope to reduce the risk of harm for the many UK children that use them. These measures also reflect that additional steps are needed for risk to be managed effectively within the context of more complex services and larger organisations, who have greater capacity to implement more costly measures. For example, these measures entail having an internal

monitoring and assurance function, and taking steps to feed negative sentiment expressed by users back into their recommender feeds.

14.57 Our proposed definition of 'large' is the same as that proposed in our Illegal Harms Consultation, capturing services with a number of monthly UK users that exceeds 7 million, which is roughly 10% of the UK population. This closely mirrors the definition of large services taken by the EU in the DSA[48] and is a threshold we have also proposed in our categorisation advice to the Secretary of State.[49] We consider it important to broadly align our approach to determining larger services with other international regimes where possible, to reduce the potential burden of regulatory compliance for services.

14.58 Consistent with our Illegal Harms Consultation, at this stage we propose that the number of monthly UK users should be calculated as an average over 12 months.[50] We will continue to consider the specific approach to user measurement in our ongoing work across the Illegal Content and Children's Codes, also having regard to the approach to user measurement in relation to categorised services thresholds, as may be specified in any future secondary legislation.

14.59 As acknowledged in our Illegal Harms Consultation, we recognise that the size of the UK user base is an imperfect proxy for a service's capacity, including its resources and capabilities. We have considered supplementing this with additional criteria but we provisionally believe that these additional criteria would still be subject to important limitations, whilst adding additional complexity, so we are not proposing these at this stage. As we explain in more detail in our Illegal Harms Consultation,[51] alternative metrics such as profit, revenue and number of employees could act as a proxy for a service's access to financial and technical resources. However, we consider that online services may have access to substantial capital even at a time when revenue, profit and employee numbers are low, for example if the user base is large or growing rapidly and funding has been raised on the expectation of greater monetisation in the future. For multinational services, there are also challenges in attributing revenue and profit to the UK market.

14.60 Our proposed definition of a large service captures services with the widest reach among UK children. Nevertheless, we recognise that the size of the total UK user base is not a precise proxy for the number of children using a service, which services are generally less able to measure accurately and robustly. In Volume 4 we discuss the relevance of the user base, including the number of children, in relation to services' risk assessments and make recommendations as to how services should take this into account.

14.61 As our understanding of costs and benefits grow, it may be proportionate in future to expand the range of services for which some measures are recommended.

---

[48] The DSA classifies platforms or search engines as very large online platforms (VLOPs) or very large online search engines (VLOSEs) if they have more than 45 million users per month in the EU, a number equivalent to 10% of the EU population.

[49] A small proportion of services in scope of the Online Safety Act will be categorised and designated as category 1, 2A or 2B services if they meet certain thresholds set out in secondary legislation by Government. Ofcom has a duty to advise the Secretary of State on threshold conditions for each category of service. See Ofcom, March 2024, Categorisation – Advice submitted to the Secretary of State.

[50] See sub-section 'User Numbers' in Annexes A7 and A8.

[51] See paragraphs 11.55 – 11.60 in Volume 4 in our 2023 Illegal Harms Consultation.

Combined impact assessment of draft Children's Safety Codes

14.62 While our impact assessment framework considers each measure individually, in practice services will be applying several measures from the Children's Safety Codes, depending on the risks they pose to children, their size and other characteristics. To supplement our individual assessments and give due consideration to the proportionality of the proposed package of measures overall, we also include a combined impact assessment of the measures in Section 23.

14.63 This combined assessment differentiates between measures based on the kinds of services they apply to, including measures which may apply to smaller services.

14.64 As explained further in Section 23, we do recommend a wide range of measures regardless of the size of service and we recognise that the cost of these measures may be high relative to the resources available to small or micro businesses. Some small or micro businesses could struggle to implement the measures, potentially leading to degradation of user experience or even withdrawal of services from the UK, potentially harming users (whether children or adults) who benefit from those services. To mitigate this risk, our measures allow for a degree of flexibility in their implementation, with costs often expected to scale with the potential benefit, in terms of reduced harm to children. Overall, we provisionally conclude that the package of measures is proportionate given its expected contribution to child safety online.

Process to implementation

14.65 Once the consultation period closes, we will consider and take into account responses and evidence received in order to prepare the final regulatory documents. This includes evidence provided in response to this consultation, as well as evidence provided in response to our Illegal Harms Consultation and consultation on the Part 5 guidance where this is relevant to the Children's Safety Codes.

14.66 We will publish a Statement on our regulatory documents and conclusions on our guidance Codes of Practice. At this point, Ofcom must submit our final draft Codes of Practice to the Secretary of State, who may set out further requirements (directions) for Ofcom in relation to our Codes where there are exceptional reasons relating to public health, national security, public safety, or relations with a government outside the United Kingdom. Otherwise, the Codes will be laid in Parliament. Unless either House of Parliament resolves not to approve the Codes within 40 days of them being laid, Ofcom will issue the Codes and they will come into force, along with the children's safety duties, 21 days later.

14.67 Any updates to the Codes, other than minor amendments, will follow a similar procedure to that set out above.

14.68 In line with our proposals in relation to enforcement under the Act more generally, our focus in the early regulatory period will be on working with services to help them understand their obligations and any steps that are needed for them to come into compliance. As protecting children online is our number one priority, this approach will be balanced against the need to take swift action against intentional or systemic breaches, and the importance of protecting children from significant ongoing harm.

14.69 We recognise that when the children's safety duties come into effect (following Codes of Practice being published), it may take time for services to bring themselves fully into

compliance. However, we expect services to take proactive steps to effectively implement safety measures as soon as is reasonably possible to protect children.

14.70 While we will consider what is reasonable on a case-by-case basis, all services should expect to be held to full compliance shortly after the relevant duty coming into effect. This means that we expect Ofcom's enforcement action to increase over time as the regime comes into effect. As protecting children online is a key priority, we will not hesitate to take swift action in relation to non-compliance with children's safety duties – for example we may wish to take early action in relation to non-compliance with age assurance duties by pornography services. We discuss our enforcement approach to all Ofcom Codes of Practice in more detail in Chapter 29 of the 2023 Illegal Harms Consultation[52].

Structure for the rest of this volume

14.71 In this volume, we set out the measures we are proposing to include in the Children's Safety Codes. These are grouped and ordered into the following sections:

15. Age Assurance
16. Content Moderation for U2U services
17. Search moderation
18. User reporting and complaints
19. Terms of service and publicly available statements
20. Recommender systems on U2U services
21. User support
22. Search features, functionalities and user support
23. Combined impact assessment
24. Statutory tests

[52] Ofcom, 2023: Protecting people from illegal harms online

15. Age assurance measures

Services that pose risk to children need to know which of their users are children to ensure they receive the protections required by the Act. Measures to establish users' ages are normally part of broader systems and processes designed to ensure children have age-appropriate experiences online. Such systems work by ensuring children are not exposed to harmful content or functionalities and are given the appropriate controls and support.

Establishing users' ages to give children the right protections should not generally result in services using those methods to deny children the benefits of being online and enjoying the opportunities that services present. The exception is where the main purpose of a service is to provide content harmful to children, and where there are no other feasible ways of managing risks to children other than excluding them.

The overarching aim of age assurance measures for services under the children's safety duties is to help ensure children are protected from harm and receive age-appropriate experiences. We have also aimed for alignment with Part 5 guidance to create clear and consistent regulatory regime for services.

Our proposals

Our proposals reflect the areas where we believe age assurance can have the most impact on the safety of children online in line with the Act requirements. The proposals are designed to help prevent and/or protect children from encountering harmful content, and to strengthen the effectiveness of other measures we are proposing in the Codes. This should help services to secure compliance with the children's safety duties and make children's experiences more age-appropriate.

Given the risk of harm to children on services in scope of our proposals, we are proposing that services should implement **highly effective age assurance** (HEAA) under the age assurance measures. Annex 10 (draft HEAA guidance) provides additional guidance on how services should interpret this term as well as examples of what methods of age assurance may be implemented in a highly effective way. In developing these positions, we have considered the current state of technology for establishing the age of users, as well as the rapid pace of development in this industry.

Establishing that a user is a child allows the service to target safety measures to them to provide appropriate layers of protection from harm. Our proposals focus on the use of age assurance as a facilitator for three types of safety measures:

- Access controls which limit children's access to an entire service or part of a service;
- Content controls which prevent or protect children from encountering harmful content; and
- Recommender systems measures which protect children from being recommended harmful content.

We recognise that our proposed recommendations on age assurance could have a potentially significant impact on the rights of users (including both children's and adults'), particularly rights to freedom of expression and privacy rights. We have therefore considered whether the degree of interference with these rights is proportionate and set out our reasoning in the detail of each measure below. We also acknowledge that our proposed

measures may be costly for services to implement, and we explain in this section why we believe our approach is proportionate given the importance of age assurance in ensuring that children have age-appropriate experiences online.

	Proposed measure	Who should implement this[53]
Service-wide access control measures		
AA1	Use **HEAA to prevent children from accessing the entire service**	All U2U services whose principal purpose is the hosting or the dissemination of one or more kinds of PPC
AA2		All U2U services • whose principal purpose is the hosting or the dissemination of one or more kinds of PC; AND • who are high/medium risk for one or more of those kinds of PC
Content control measures		
AA3	Use **HEAA to ensure children are prevented from encountering PPC identified** on the service	All U2U services • whose principal purpose is not the hosting or the dissemination of one or more kinds of PPC; AND • which do not prohibit one or more kinds of PPC
AA4	Use **HEAA to ensure children are protected from encountering PC identified** on the service	All U2U services • whose principal purpose is not the hosting or the dissemination of one or more kinds of PC; AND • which do not prohibit one or more kinds of PC; AND • are high/medium risk for one or more kinds of PC that they do not prohibit
Targeting recommender systems measures		
AA5	Use **HEAA to apply relevant recommender system measures** in the Code **to children**	All U2U services that • are high/medium risk for one or more kinds of PPC; AND • operate a content recommender system
AA6		All U2U services that • are high/medium risk for one or more kinds of relevant PC (excluding bullying); AND • operate a content recommender system

Consultation questions

31. Do you agree with our proposal to recommend the use of highly effective age assurance to support Measures AA1-6? Are there any cases in which HEAA may not be appropriate and proportionate? In this case, are there alternative approaches to age assurance which would be better suited? Please provide any information or evidence to support your views.
32. Do you agree with the scope of the services captured by AA1-6?
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?

[53] These proposed measures relate to providers of services likely to be accessed by children.

34. Do you have any comments on our assessment of the implications of the proposed Measures AA1-6 on children, adults or services? Please provide any supporting information or evidence in support of your views.

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?

What is age assurance?

15.1 Age assurance measures can be used to ensure the online experiences of children are safe while preserving the rights of adult users to access legal content. By distinguishing between children and adult users through age assurance, services can provide age-appropriate experiences to their users.

15.2 Determining age is routinely used offline as an important component of ensuring children's safety. This includes when preventing children from buying restricted goods (e.g., alcohol) or accessing restricted areas. The same is not true in the online world, where children can often access adult experiences without any restrictions, exposing them to age-inappropriate experiences that can result in harm.

15.3 The online age assurance industry is developing rapidly. It is likely to continue to grow as the demand on age assurance providers and service providers to offer users the best experience increases. We have already seen, for instance, large service providers starting to develop their own proprietary methods in-house alongside an array of offers from third party age assurance providers. As governments, service providers and consumers continue to prioritise online safety for children, we expect barriers to engaging with this technology to reduce and enable the development of age assurance methods to accelerate. In the United Kingdom, age assurance is already used in other regulated online sectors. For example, the law requires all online gambling businesses to ensure that users provide a form of identification for age, financial and identity verification.[54]

15.4 Our draft Children's Safety Codes contain a full package of proposed measures to secure safer and more age-appropriate children's experiences online. This section sets out the detail of our proposed measures and reasons for recommending them, establishing our expectations on services to determine which users are, or are not, children to target their safety measures effectively. This is the first step in ensuring the wider efficacy of the safety measures contained in the draft Children's Safety Codes, such as those related to recommender systems.

15.5 Age assurance, together with content moderation, service design and user support, should work to disrupt the ease with which children are currently exposed to, and can access harmful content, as well as the prevalence and dissemination of such content.

[54] Gambling Commission, 2024. Age, ID and financial verification. [accessed 24 April 2024].

Definition Box 1: Relevant terms

Access controls: mechanisms to determine which users can access online content or spaces.

Age assurance: a collective term for age verification and age estimation.

Age estimation: a form of age assurance designed to estimate the age or age-range of the user[55], for example using facial age estimation.

Age verification: a form of age assurance designed to verify the exact age of the user[56], for example using a form of identity documentation.

Age check: An individual instance where a user is required to undergo an age assurance process.

Content controls: mechanisms to determine the visibility and accessibility of content including its removal or reduction.

Highly effective age assurance: methods of age assurance that are of such a kind and implemented in such a way that is highly effective at correctly determining whether or not a particular user is a child.

Self-declaration: a process where the user is asked to provide their own age. This could be in the form of providing a date of birth to gain entry to a service or by ticking a box to confirm a user is over a minimum age threshold.

Our proposals to protect children

15.6 The Act states that U2U services likely to be accessed by children must use proportionate systems and processes designed to prevent children from encountering PPC on the service, and to protect children in age groups judged to be at risk of harm from encountering PC and NDC that is harmful to children.[57]

15.7 Section 12(4) of the Act requires service providers to use age assurance (which must be highly effective) to prevent children encountering PPC that the service provider identifies on its service. Section 12(5) of the Act clarifies that such a requirement applies to a provider in relation to a particular kind of PPC in every case, except where the terms of service prohibit that kind of PPC on the service and that policy applies to all users of the service. While the Act does not specifically require age assurance to be used for Part 3 services in other scenarios, it is listed as a measure that may be taken or used (among others) to comply with the duties in section 12(3).[58]

15.8 Regulated search services must use proportionate systems and processes designed to minimise the risk that children are exposed to PPC and PC.[59] We are not proposing to recommend the use of age assurance for search services as we believe the measures proposed in relation to Sections 17 (search moderation) and 22 (search features,

[55] Section 230(3) of the Act.
[56] Section 230(2) of the Act.
[57] Section 12(3) of the Act.
[58] Section 12(7) of the Act.
[59] Section 29(3) of the Act.

functionalities and user support) can achieve these outcomes without the use of highly effective age assurance. We discuss the rationale for this further in Sections 16 and 22.

15.9 For this reason, in the remainder of this section, references to 'services' and 'providers' refer only to U2U services and service providers likely to be accessed by children.

15.10 In developing our recommendations, we have considered the principles set out in paragraph 12(2), Schedule 4 to the Act. This includes taking into account the nature and severity of potential harm to children in line with the principle that "more effective kinds of age assurance should be used to deal with higher levels of risk of harm to children."[60] In doing so, we have also considered the cost impact on businesses, the potential impact on children and adult users, and the need to protect their rights to privacy and freedom of expression. In addition, in developing our proposals on kinds of age assurance we have sought to ensure accessibility, including by children, and effectiveness for all users regardless of their characteristics. Where possible, we also considered the principle of interoperability between different kinds of age assurance. We have explained how we have taken these impacts and principles into account where relevant in outlining our proposals, and particularly in deciding how the level of risk posed by a service influences the circumstances in which the use of age assurance may be needed.

15.11 We are not recommending the use of specific age assurance methods in our measures. We have instead recommended that, to ensure that their age assurance process is highly effective, services take steps to fulfil the criteria of technical accuracy, robustness, reliability and fairness. This flexibility will enable services to choose their approach to meeting these criteria in a way that is most cost-effective and technically feasible for them. We provide further guidance on the implementation of highly effective age assurance in Annex 10.

15.12 Children deserve the same level of protections when it comes to pornographic content, regardless of the type of service that offers it. We have therefore considered the need for consistency in our approach to age assurance for pornographic content, whether this is in the context of preventing children from accessing pornographic content, a form of PPC, for the purposes of the Part 3 children's safety duties or under the obligations set out in Part 5. Under Part 5 of the Act, service providers who publish regulated provider pornographic content are required to implement highly effective age assurance to ensure that children are not normally able to encounter such content on their service.[61] It is important that we set consistent expectations for how service providers that allow pornographic content on their service implement highly effective age assurance to prevent children from encountering pornographic content, regardless of whether the Part 3 and/or Part 5 duties apply. This will ensure that children experience the same level of protection on all services. Our proposed approach to highly effective age assurance for Part 3 therefore mirrors the approach set out in our draft guidance for service providers publishing pornographic content.[62]

15.13 We are in the process of analysing responses to our Part 5 Consultation on the draft guidance for service providers publishing pornographic content. In finalising our proposed approach to the implementation of highly effective age assurance we will consider

[60] Paragraph 12(2)(d) of Schedule 4 to the Act.
[61] 'Regulated provider pornographic content' is pornographic content which is published or displayed on an online service by the provider of the service, or by a person acting on behalf of the provider, as set out in Section 79(2) of the Act.
[62] Ofcom, 2023. Guidance on age assurance and other Part 5 duties for service providers publishing pornographic content on online services. Annex 2.

stakeholder comments in response to that consultation, alongside comments we receive in response to our age assurance proposals in this consultation.

15.14 All methods of age assurance involve the processing of personal data. When implementing age assurance, service providers will be expected to comply with data protection laws and in particular, to have regard to the provisions of the ICO's Children's code on the processing of children's personal data. The ICO have also published an Information Commissioner's Opinion about the use of age assurance for their Children's code which will be relevant to service providers in scope of this section.[63]

Our proposed measures

15.15 We are proposing to recommend the following measures for U2U services. We set out an explanation of how each measure works in detail below:

A) Service-wide access control measures

- **Measure AA1:** Services whose principal purpose is the hosting or the dissemination of one or more kinds of PPC should use highly effective age assurance to prevent children from accessing the entire service.
- **Measure AA2**: Services whose principal purpose is the hosting or the dissemination of one or more kinds of PC, and who are high or medium risk for one or more of those kinds of PC, should use highly effective age assurance to prevent children from accessing the entire service.

B) Content control measures

15.16 **Measure AA3:** Services whose principal purpose is not the hosting or the dissemination of one or more kinds of PPC, and which do not prohibit one or more kinds of PPC, should use highly effective age assurance to ensure children are prevented from encountering PPC identified on the service.

- **Measure AA4:** Services whose principal purpose is not the hosting or the dissemination of one or more kinds of PC; and which do not prohibit one or more kinds of PC; **and** are high or medium risk for one or more kinds of PC that they do not prohibit should use highly effective age assurance to ensure that children are protected from encountering PC identified on the service.

C) Targeting recommender systems measures

- **Measure AA5:** Services that are high or medium risk for one or more kinds of PPC and operate a recommender system, should use highly effective age assurance to apply the relevant recommender system measures in the Code to children.
- **Measure AA6:** Services that are high or medium risk for one or more kinds of relevant PC and operate a recommender system, should use highly effective age assurance to apply the relevant recommender system measures in the Code to children.

15.17 We assess the impact of these measures below and explain why our provisional view is that they would be proportionate interventions.

[63] ICO, 2024. ICO's updated opinion on the Children's Code. [accessed 18 April 2024].

15.18 In determining who our measures should apply to, we initially considered focusing solely on the outcome of the children's risk assessment, e.g., recommending highly effective age assurance only on the basis that a service was high or medium risk for any kind of PPC or PC appearing. However, we considered that these measures would be too broad, as they would not target the areas where access controls and content controls are likely to have the most impact. Level of risk remains an essential component in our proposed measures, and a service's children's risk assessment will be an important tool for service providers to determine what risk of content harmful to children they have on their service. In addition, our proposals recognise that factors such as whether a service prohibits harmful content; hosts or disseminates harmful content as its principal purpose; or, has functionalities that amplify the risk of encountering harmful content such as a recommender system, all of which play an important role in how children are exposed to harmful content.

15.19 In the discussion of our proposed measures, 'users who have not been determined to be adults' refers to users whose age has not been established to be 18+ by means of highly effective age assurance. While the Act recognises the need to protect children in different age groups judged to be at risk of harm from encountering PC and NDC, we are not proposing the use of age assurance to determine the specific age groups of users below the age of 18. The reasons for this are stipulated in the 'Children in different Age Groups' sub-section of this section. We may look to adjust our recommendations on PC to focus on specific age groups in the future, as technology evolves and depending on the responses to this consultation. In the meantime, the measures proposed in this section are intended to be a complement to, rather than a substitute for, measures that services may apply to tailor their services to offer age-appropriate experiences for children of different ages.

The role of age assurance for other protection of children measures

15.20 Where service providers implement highly effective age assurance in accordance with any of the age assurance measures proposed in this section, this may be used to target other relevant safety measures to children. This includes the following safety measures recommended in Section 21, as outlined below:

- Measure US1 - Providing children with an option to accept or decline before being added to groups.
- Measure US2 - Providing children with the option to block or mute other user accounts on the service.
- Measure US3 - Providing children with the option of disabling comments on their own posts.
- Measure US4 - Prompting children when they take action to restrict their interactions with another user or a particular type of content with information about how to report harmful content.
- Measure US5 - Signposting children to support when they:
    - report bullying, suicide, self-harm or eating disorder content;
    - post or share bullying, suicide, self-harm or eating disorder content; or,
    - search for suicide, self-harm or eating disorder content.

15.21 If a service provider does not wish to use highly effective age assurance to target these safety measures at children specifically, it should instead apply those safety measures to all users. This recommendation is set out in Section 21.

15.22 Similarly, we are not proposing to recommend age assurance for search services for reasons discussed in Sections 17 and 22. Should a service provider choose to implement highly

effective age assurance and target measures exclusively towards children, then they may do so rather than implementing our search measures to all users.

Age assurance for measures to prevent children from encountering illegal harms

15.23 In our Illegal Harms Consultation, we set out a package of measures relating to the default settings of child user accounts on U2U services, and the provision of supportive information at critical points of a child user's online experience. For instance, if a service provides the relevant functionality, it should ensure that children are not presented with prompts to expand their network of friends or included in network expansion prompts presented to others. We proposed that services should provide children with supportive information when they are seeking to disable one of the default settings recommended; responding to a request from another user to establish a formal connection; exchanging a direct message with another user for the first time; and, taking action against another user, including blocking and reporting.[64] The proposed package of measures aims to mitigate risks to children encountering illegal harm, with a specific focus on grooming for the purposes of sexual abuse.

15.24 The proposed measures in the Illegal Harms Consultation rely on services having an existing means of identifying whether users are children and would apply where the information available to services indicates that a user is a child. In our Illegal Harms Consultation, we anticipated that, where services are already using age assurance technologies, they would use these to determine whether a user is a child for these purposes.

15.25 When service providers in scope of the age assurance measures proposed in this consultation determine a user is a child through highly effective age assurance, services should use this information to target the proposed Illegal Harms Consultation measures on default setting and supporting information.

15.26 The recommendations in this section would strengthen the effectiveness of the default settings and supportive messaging measures in our Illegal Harms Consultation by ensuring that the service is better able to determine which users are children so they can benefit from the protections that these measures offer.

15.27 We have received responses to our Illegal Harms consultation and are currently considering these responses, which we will do jointly with responses to this consultation.

Current practices

15.28 This section sets out what we know of existing age assurance processes used by different services and the risks posed to children by ineffective age assurance. Currently, most services either do not use age assurance or use processes that are not effective.

15.29 Self-declaration is widely used as a method by tech providers as a means to limit access to restricted content, and to gather information for the purpose of targeting experiences to users. The Act explicitly states that self-declaration is not a form of age assurance.[65] Evidence suggests that self-declaration is an ineffective method for establishing the age of a user. Ofcom research found that a fifth of children aged eight to twelve with a social media profile have a user age of 18 or over on at least one service. Even if children do not have an

[64] Ofcom, 2023. Consultation: protecting people from illegal harms online, Volume 4, Section 18.
[65] Section 230(4) of the Act.

adult pr just under two thirds (64%) of children aged eight to twelve with a social media profile have a user age of 13-15.[66]

15.30 Even in the case of services specifically targeted towards adults, self-declaration is widely used as the only measure for controlling access. For example, British Board of Film Classification (BBFC) research found that 63% of the top 100 most accessed pornography services did not have any measures in place for identifying the age of the user and of the remainder that do, self-declaration is the most used method.[67]

15.31 Ofcom's Video Sharing Platform (VSP) report into how VSPs protect children found that TikTok, Twitch and Snap rely on self-declaration as the first step in establishing the age of their users. They then implement additional measures after account creation to try to identify underage (under 13) accounts.[68] The services use a range of methods to attempt to detect underage users, including keyword detection, user reporting and flagging and analytical tools. Additionally, to detect child users generally (under 18), the report found that Twitch used human moderators and Snap relied on an inferred age model.

15.32 In response to our 2023 Protection of Children Call for Evidence (our 2023 CFE), Match Group told Ofcom that they use self-declaration at registration followed by artificial intelligence content moderation and human moderation to identify signs of users under 18 once registered, for example through pictures, bios and conversations.[69]

15.33 While services may use other methods alongside self-declaration to establish the age of their users, this ex-post approach allows for children to access the service and potentially encounter harmful content before they are later found to be children. There is no independently verified information on the effectiveness of these complementary methods in helping to validate the age of the user.

15.34 In our 2023 CFE, X – formerly Twitter– also said it relied on self-declaration at account creation. Users whose date of birth placed them as over the age of 13 but under the age of 18 had additional safety measures on their account, which the user could choose to turn off.[70]

15.35 In response to our 2023 CFE, Google stated that, across their consumer facing products, in addition to self-declaration, they use a machine learning model to "help infer if a user is over or under the age of 18 based on a variety of behavioural signals."[71] During account creation, users are prompted to provide their birth date. If a user tells Google they are under the age of 13, the service directs the user to the Family Link account creation flow to create a supervised account. If a user tells Google they are over 13 years old but below 18, Google offers them default protections. The machine learning tool is deployed to provide an additional level of assurance of a user's declared age, or to indicate whether or not a user is a child where they have said they are over 18 or have not declared their age (e.g., accessing the service in a logged-out state). Google might require age verification if a user is trying to

---

[66] Ofcom, 2024. Children's Online User Ages 2024 Quantitative Research Study.
[67] Of the top 100 most accessed pornography services by the UK, the British Board of Film Classification (BBFC) found that 37% required some form of age check. In all but one case this relied on a user's self-declaration of age. Research conducted between August 2022 and March 2023. Ofcom, 2024. Functionality of Online Pornography Services: A BBFC research report for Ofcom.
[68] Ofcom, 2023. How video-sharing platforms (VSPs) protect children from encountering harmful videos.
[69] Match Group response to 2023 Protection of Children Call for Evidence.
[70] Twitter response to 2023 Protection of Children Call for Evidence.
[71] Google response to 2023 Protection of Children Call for Evidence.

access age-restricted content and the service cannot establish with sufficient certainty that a user is an adult; or the age assurance process has estimated the user as under 18 and they wish to access age-restricted content. Users can either verify their age using a government ID or a credit card or, in some jurisdictions including the UK, use a selfie for age estimation.[72]

15.36 Pinterest uses self-declaration to enforce its minimum age of 13 for account creation. It places the highest privacy settings on the accounts of users whose self-declared age places them between 13 and 17 by default.[73] In addition, Pinterest applies safe messaging features to these users, which includes ensuring that they can only receive messages from users who are known to them. Where Pinterest subsequently finds a user to be under the age of 13, either due to self-declaration or a report by their parent, it deletes that account. Users can use age verification provided by a third-party provider to appeal this decision.[74]

15.37 Some services designed for children have implemented measures to understand user characteristics, including age. Lego's Verified Parental Consent asks for either ID or credit card verification to allow parents to sign up for an account and link it to their child's account. Here, only the parent must undergo age assurance, not the child.

15.38 Yubo is a service targeted at children and young people. It uses Yoti, a third-party solution based on facial age estimation to check user age, supported by identity verification where an age estimation result requires additional checks.[75]

What harms do age assurance measures protect children from?

15.39 Our evidence suggests that many children in the UK encounter harmful content online, and that the impacts of viewing harmful content are wide-ranging and can be severe. We have documented the extensive impacts that encountering PPC and PC have on children, which in severe cases can lead to death, for more information see Volume 3 of this consultation.

15.40 As discussed under current practices above, services are not typically using effective processes to age assure their users. Without age assurance, services cannot apply safety measures targeted at children in a way that ensures that all child users will benefit from an appropriately tailored experience. Our research on pathways to violent content and experiences of cyber-bullying among children indicates some children support approaches to age assurance that go further than self-declaration, to protect them from content or functionalities intended for adults. Children were supportive of more accurate approaches to age assurance and provide suggestions such as linking ID to national insurance numbers.[76]

15.41 Our proposed measures seek to address this and minimise the likelihood and impact of exposure to harmful content outlined in Volume 3.

15.42 In developing our measures, we considered evidence that harmful content can still be widespread where services prohibit it in their terms of service as per the overview of the codes in Section 13.

---

[72] Google, 2024. Access age-restricted content and features. [accessed 12 March 2024].
[73] Pinterest response to 2023 Protection of Children Call for Evidence.
[74] Pinterest response to 2023 Protection of Children Call for Evidence.
[75] Yubo, Safety Tools. [accessed 3 January 2024].
[76] Across our research programme, children shared their experiences and told us what they want and need to be protected online. In two studies children demonstrated support for more accurate age assurance: Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children and Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.

15.43 As well as this, our evidence indicates that children are at a higher risk of encountering harmful content, including sexual content, violent content and suicide and self-harm content, on services which deploy recommender systems.[77] The way in which these mechanisms enable services to push harmful content to children means that recommender systems have also been an important factor in the development of our age assurance measures.

15.44 Our proposals recognise that age assurance is not a silver bullet, and it will not be the most effective way to protect children in all scenarios. For instance, where a service prohibits all PPC and PC but this content is readily available in the service, the provider should focus on improving its content moderation systems and reviewing its wider risk management processes as per Section 16 and our draft Children's Risk Assessment Guidance (Annex 6). Accordingly, our content control measures (see Measures AA3 and AA4 below) apply only where services do not prohibit one or more kinds of PPC and PC. These services will need to take additional steps to secure that children are appropriately protected from these kinds of content that they choose to allow adult users to access on the service.

Service-wide access control measures

15.45 Access control measures are mechanisms to determine which users can access online services. Before users can access services that host harmful content, they may be prompted to go through an age assurance process. Our provisional view is that age assurance is essential in facilitating effective access control measures for services whose principal purpose is the hosting or dissemination of PPC or PC, and that would realistically have no other way to prevent children from encountering this content other than to prevent them from accessing the service. Measures AA1 and AA2 therefore recommend highly effective age assurance to support the use of service-wide access controls.

Measure AA1: Use HEAA to prevent children accessing services whose principal purpose is the hosting or dissemination of PPC

Services whose principal purpose is the hosting or the dissemination of one or more kinds of PPC should use highly effective age assurance to prevent children from accessing the entire service.

Explanation of the measure

15.46 Services in scope of this proposed measure should use effective access controls to prevent users from accessing the service unless they have been determined to be adults.

15.47 This measure applies to services whose principal purpose is to host or disseminate one or more kinds of PPC. 'Principal purpose' in this context refers to the main activity or objective of the service.

[77] See Section 7, 7.1, 7.2, 7.6. Detailed explanations on how recommender systems work and how they can pose a risk to children is set out in recommender systems on U2U services Section 19.

15.48 It is for the service provider to assess the nature and purpose of its site to determine whether its principal purpose is the hosting or dissemination of PPC. Relevant indicators could include, but are not limited to:

- Whether the service promotes or refers to any kind of PPC, for instance through its name, branding, terms of service, or any other means of describing the service to users; and how it markets or positions itself against its competitors;
- How the content itself is presented or described, including consideration around whether PPC is the main draw for users of the service;
- Whether it provides access to content other than PPC. In cases where it does, it may be relevant to consider the centrality of PPC to the service, including the proportion or relative prominence of PPC on the service. We would expect a service's principal purpose to be the hosting or dissemination of PPC where the content present on the service, taken overall, is entirely or predominantly comprised of PPC.

15.49 In making this determination, services may find it helpful to consult our Guidance on Content Harmful to Children in Volume 3, Section 8 which provides examples of content which Ofcom would, and would not, consider to be PPC.

15.50 We would expect services in this category to include dedicated pornographic services and certain discussion forums or chat rooms where suicide, self-harm and eating disorders are the primary subjects of discussion. This list is non-exhaustive, and we welcome evidence from stakeholders on other types of service whose principal purpose is to host or disseminate PPC.

15.51 If a proportion of service's residual content is not considered harmful to children (i.e. it is not PPC/PC or NDC), services may wish to consider whether they can create a child-safe experience on their service. We would consider that this is unlikely to be possible if only a minority of residual content is not harmful to children.

15.52 The service provider should implement effective access controls to prevent users from accessing the service unless they have been determined to be adults. For the purposes of AA1 and AA2 this includes any part of the service on which regulated user-generated content is or may be present. To prevent access to the entire service, the service provider should implement highly effective age assurance and effective access controls in a way, and at a point in the sign-in process, that prevents users from encountering PPC content on the service before the service has determined their age. This means implementing age assurance at the point of entry to the site and/or ensuring that no PPC is visible to users on entering the site before they have completed an age check.

15.53 The effectiveness of an age assurance method will depend on how it is implemented, including whether by itself or in combination with other age assurance methods. For the purposes of meeting this proposed measure, the age assurance process as a whole needs to be highly effective at correctly determining whether or not a particular user is a child. We provide draft guidance on implementing highly effective age assurance in the draft HEAA guidance at Annex 10.

15.54 The effect of deploying highly effective age assurance in this way should be that it is no longer possible for children to normally access the service, and so the service will no longer be likely to be accessed by children. Services that are not likely to be accessed by children

are not in scope of the children's risk assessment and safety duties.[78] Where a service implements highly effective age assurance to prevent children from accessing the service, it will normally be out of scope of the children's safety duties in line with the principles established in relation to children's access assessments (Volume 2, Section 4).

Effectiveness at addressing risks to children

15.55 Highly effective age assurance is a requirement under the Act for services that do not prohibit particular kinds of PPC in their terms of service.[79] We expect services in scope of Measure AA1 will not prohibit one or more kinds of PPC as hosting or disseminating this content will be their principal purpose. This means that we do not have discretion to recommend the use of any form of age assurance which is less effective in these circumstances.

15.56 In addition, taking the Schedule 4 principles into account, highly effective age assurance aligns with the principle that more effective kinds of age assurance should be used to deal with higher levels of risk of harm to children.[80]

15.57 We have exercised a degree of discretion in recommending that services in scope of Measure AA1 prevent users from accessing the entire service unless they have been determined to be adults, rather than only preventing access to identified PPC as required by the Act. This is because the risk of children's exposure to PPC on services that are dedicated to this content is almost certain and there are no realistic alternative ways in which the service may be able to manage the risk of children being exposed to harmful content. For example, a tube site dedicated to the sharing of pornographic user-generated content is not likely to able to manage the risk of children being exposed to this content other than by restricting access to the service. We consider that preventing access to the entire service using effective service-wide access controls is the only feasible solution to prevent children from encountering PPC in practice.

15.58 As well as reflecting the risk of harm to children, this will help to ensure a consistent approach across our proposals on content types that fall within the PPC definition. This is relevant specifically in relation to the requirements for regulated provider pornographic content under Part 5 of the Act, and for dedicated pornography services that fall in scope of Part 3.[81] The use of effective access controls to prevent access to services by children is an

[78] Part 3 services that are likely to be accessed by children will be in scope of the children's risk assessment duties and safety duties protecting children in the Act. Services can only conclude it is not possible for children to assess a service, or part of it, if they are using age assurance with the result that children are not normally able to access the service (section 35(2) of the Act). As discussed in Section 4, we propose that service providers should only conclude that it is not possible for children to access a service, or part of the service, where they are using highly effective age assurance to secure this outcome.

[79] See sections 12(3)(a), 12(4), 12(5) and 12(6) of the Act.

[80] Paragraph 12(2)(d) of Schedule 4 to the Act.

[81] Dedicated pornography services may fall under Part 3 and/or Part 5 of the Act where they have a mix of pornographic content that is user-generated and provider pornographic content. Where the majority of the content hosted by a dedicated pornography service is user-generated pornographic content, the service should fall in scope of age assurance Measure 1. Where a service also hosts provider pornographic content under Part 5 of the Act, while that content does not fall into the definition of PPC, service providers are required to ensure that children cannot normally encounter that content through the use of highly effective age assurance. Applying Measure 1 where the principal purpose of the service is to host or disseminate pornography should therefore secure compliance with both the Part 3 and Part 5 duties requiring the use of highly effective age assurance for pornography (Section 12(4)-(6) and section 81(3) of the Act).

important element of our draft guidance for service providers publishing pornographic content under Part 5.

15.59 We are mindful of the fact that it may be possible for users to circumvent individual age assurance methods, as well as the age assurance process or access controls. We note that the benefits of this measure could be reduced if there are opportunities for children to circumvent the age assurance process. As explained in our draft codes measure relating to the implementation of highly effective age assurance, service providers should take steps to identify any methods children are likely to use to circumvent the age assurance methods implemented and take feasible and proportionate steps to mitigate against the use of these methods of circumvention, in so far as it is possible to do so.

Rights assessment

15.60 This proposed measure recommends that all services that are likely to be accessed by children and whose principal purpose is the hosting or dissemination of one or more forms of PPC use highly effective age assurance to ensure that children are prevented from accessing the service. It is designed in accordance with our criteria-based approach to implementing highly effective age assurance and does not mandate a specific method of age assurance.

15.61 This measure may have a potentially significant impact on the rights of users (including both children and adults[82]) to privacy (Article 8 of the ECHR), freedom of religion and belief (Article 9 of the ECHR), freedom of expression (Article 10 of the ECHR) and freedom of association (Article 11 of the ECHR). It may also have a potentially significant impact on service providers' rights to freedom of expression. We have therefore considered the extent to which the degree of interference with these rights is proportionate.

15.62 In considering the degree of the potential impact on users' and services providers' rights and whether it is proportionate, we have taken as our starting point the requirements of the Act. The children's safety duties set out in the Act require providers of U2U services to use proportionate systems and processes to prevent children from encountering PPC.[83] They also require services that do not prohibit all kinds of PPC to use highly effective age assurance as part of their systems and process to prevent children encountering PPC.[84] By preventing children's access to PPC, the proposed measure will seek to secure adequate protections for children from harm, in line with the legitimate aims of the Act. It also aims to secure that a higher level of protection is provided to children than adults. Preventing children from encountering PPC acts to prevent the harmful consequences that such content can have on them, including to children's physical, mental or emotional wellbeing. We therefore consider that a significant public interest exists in measures which aim to prevent children from encountering PPC. This substantial public interest relates to the protection of children's health and morals, public safety, and particularly the protection of the rights of others, namely child users of regulated services.

[82] Adult users also include those who are operating on behalf of a business, or accounts that might also be concerned with other entities, such as charities, as well as those with their own, individual account. Both corporate and individual users can benefit from the right to freedom of expression, and we acknowledge the potential risk of interference with the rights of these users to freedom of expression, in addition to the rights of children and adults as individuals.

[83] Section 12(3)(a) of the Act.

[84] Section 12(4)-(6) of the Act.

Freedom of expression and association

15.63 As explained in Volume 1, Section 2, Article 10 of the ECHR upholds the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 11 of the ECHR upholds the right to associate with others. Any interference with the right to freedom of expression and association must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest.

15.64 With this proposed measure, potential interference with both child and adult users' rights to freedom of expression and association, and service providers' rights to freedom of expression, arises where the service provider applies highly effective age assurance with the objective of restricting children's access to the entire service to prevent them encountering PPC. As noted above, the duty for services that do not prohibit all kinds of PPC to use highly effective age assurance to prevent children from encountering PPC identified by the service is a requirement of the Act. To the extent that the result of implementing the proposed measure is that children are effectively prevented from encountering PPC identified on the service, and adults (including content creators) are restricted from sharing such content with children, we consider this the minimum action required to secure that the kinds of services in scope of this measure meet their duties under the Act.

15.65 We note, however, that this proposed measure would prevent children from accessing any regulated user-generated content on the service, including any non-PPC on the service which they could benefit from, and this goes further than the children's safety duties in the Act strictly require.[85] As discussed above, services in scope of this measure will be services whose principal purpose is to host or disseminate PPC, with the content on the service consisting entirely or predominantly of PPC. Therefore, we consider that the amount of non-PPC on such services from which children could potentially benefit is likely to be very limited. For the reasons explained above, we also consider that preventing access to the entire service using effective service-wide access controls is the only feasible solution to prevent children from encountering PPC in practice. Therefore, we do not consider there is a less intrusive way for services in scope of this proposed measure to meet the requirements of the Act.

15.66 While the proposed measure does not recommend services restrict adult users' access to the service or the content on it, the implementation of this proposed measure could in some cases result in potentially significant impacts on adult users' ability to access the service. This is particularly the case in the following circumstances.

15.67 First, as we explain below, we recognise the costs of implementing this measure may be significant, such that some services may not be able to carry the cost burden of implementing age assurance, for instance, smaller services which do not prohibit one or more types of PPC, and may decide to exit the UK market. This would mean that UK adults would also no longer be able to access these services, thus having a significant impact on their rights to receive them and potentially to associate with other users through these services. However, we consider it highly unlikely that all services in scope of this measure would cease to operate in the UK. For example, we would expect that many dedicated pornography services would continue to make themselves available to UK adult users, who

[85] We note that it is possible that this could include content related to religion or belief which could engage users' rights under Article 9 of the ECHR, although we consider the likely impact in this regard to be limited, given the nature of the services we propose would be in scope of this measure.

would therefore be able to access pornographic content on those services, even if their choice of such services overall were to be more limited than it is currently.

15.68 Second, we acknowledge that our measures will make it more cumbersome for adults to access these services, and the way services implement age assurance could in some cases dissuade adult users from using the service altogether. For example, some services may make their service only available to users with accounts, to reduce costs by requiring a one-off age assurance check. This may result in reduced ability for adults to access the service without being logged in, which could also have an adverse impact on their rights to receive information via these services and potentially to associate with others on these services if they would be dissuaded from accessing them as a result (for example, due to concerns about how their personal data might be used if they have to create an account on such services or how their activity may be tracked by the service). Where services choose to implement the age assurance process so that adult users are not required to create an account to access the service, we acknowledge it is also possible that some adult users might prefer not to complete age assurance each time they seek to access the service as they may find this onerous, or due to privacy concerns, and therefore may be dissuaded from using the service as a result. We consider this impact on their freedom of expression and association rights to be relatively limited, given they will have a viable option to access the service and the content on it if they assure their age, and it would therefore be their choice not to use this mechanism. In both cases, we consider these risks will also be potentially limited by the fact that providers have incentives to make their age assurance process as user-friendly as possible and limit friction to adult users. This would also limit the risk that some adults may find it more difficult to assure their age under certain methods, e.g. if they do not have the required documentation to confirm they are an adult. We have also reflected the importance of this via our proposed approach to implementing highly effective age assurance, in that we propose services should consider the principle that age assurance should be easy to use.[86]

15.69 Finally, we note that some adult users may be inadvertently restricted from accessing the service because the age assurance process assesses them to be a child. While there is potential risk for a margin of error in the use of highly effective age assurance, we consider this risk to be limited provided that services take account of our recommendations at 'Our approach to highly effective age assurance' and Annex 10 (draft HEAA guidance) to ensure the age assurance method implemented is done so in a way that is highly effective. See also the discussion of privacy impacts below and the relevance of data protection requirements which may also mitigate the impact on the adult user's rights to freedom of expression and freedom of association by giving the user a mechanism for redress and providing a route to rectify negative impacts by allowing adult users access to the service.

15.70 While we recognise the potentially significant impacts on users' rights to freedom of expression and association, as outlined above, the proposed measure is likely to go no further than needed to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the significant benefits to children into consideration, we consider that the interference with users' rights to freedom of expression and association is therefore proportionate.

[86] Age assurance should be easy to use and work for all users, regardless of their characteristics or whether they are members of a certain group. Please refer to our draft HEAA guidance at Annex 10 for the practical steps for services to consider.

15.71 The proposed measure may also have an impact on service providers' rights to freedom of expression, in particular their right to impart information to users in the UK. This would particularly be the case if, as a result of introducing highly effective age assurance, UK adult users were dissuaded from using these services, or if they were to cease to make themselves available to users in the UK due to the cost burden involved in implementing the measure. However, we consider that most of this impact arises from the duties placed on service providers under the Act, rather than as a result of the way that Ofcom is proposing they comply with these duties, as for the reasons outlined above, we do not consider there is a less intrusive way for services in scope of this proposed measure to meet the requirements of the Act. For the above reasons, and taking into consideration the significant benefits to children from preventing the harmful consequences of their exposure to PPC on these services which may otherwise occur, our provisional view is that the impact service providers' rights to freedom of expression is therefore proportionate.

Privacy

15.72 As explained in Volume 1, Section 2, Article 8 of the ECHR confers the right to respect for individuals' private and family life. Any interference with the right to privacy must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest. Again, to be 'necessary', the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.

15.73 All methods of age assurance will inevitably involve the processing of personal data of individuals, including children, whose personal data requires special consideration.[87] It will therefore impact on users' rights to privacy and their rights under data protection law. The degree of interference will depend on the extent to which the nature of their affected content and communications is public or private, or, in other words, gives rise to a legitimate expectation of privacy. It will also depend on the nature of the information required to complete the highly effective age assurance process, for example, the more sensitive information required, the more intrusive the method of highly effective age assurance is likely to be.

15.74 This proposed measure is not limited only to content or communications that are communicated publicly[88], and may lead to impacts on children's – and for the reasons noted above, adults' – ability to access services. To the extent that a service in scope of this measure may provide means for users to communicate privately (e.g. private messaging functionalities) or communications in relation to which individuals might expect a reasonable degree of privacy, this would in turn lead to more significant privacy impacts than in connection with impacts on content or communications that are widely publicly available (whether on the service concerned or more generally). We note that some of these impacts may be unavoidable: for example, preventing children from accessing any means of using the service for the purposes of private communications. Other impacts on adults'

[87] Per Recital 38 of the UK GDPR.

[88] As part of its consultation on illegal harms Ofcom consulted on draft guidance on content communicated 'publicly' and 'privately' under the Online Safety Act. That guidance recognises that whether content is communicated 'publicly' or 'privately' for the purposes of the Act will not necessarily align with whether that content engages users' (or other individuals') rights to privacy under Article 8 of the European Convention on Human Rights. For example, it is possible that users might have a right to privacy under Article 8 of the ECHR in relation to content which is communicated 'publicly' for the purposes of the Act. Conversely, users may not have a right to privacy under Article 8 of the ECHR in relation to content which is nevertheless communicated 'privately' for the purposes of the Act.

rights, as outlined above, would be more limited to the extent that access to services in scope of these measures continues to be available to them provided they assure their age. However, we acknowledge that for services that offer these functions, the proposed measure may still have some impact on adult users, depending on the way that services choose to implement the measure.

15.75 We have considered carefully whether we should limit this measure such that it does not apply to private communications and/or content communicated privately so as to limit these potentially significant impacts on users' rights to privacy. We do not consider this to be appropriate because the nature of the services in scope of this proposed measure means that we do not consider that it is likely that children could be prevented from exposure to PPC in any private communications functionalities enabled by these services, given the principal purpose of the service would be to host content related to one or more forms of PPC. We also consider that any functionalities enabling private communications are likely to be an ancillary function of such services, as we would anticipate that they will generally be focused on open communications – for example, services dedicated to pornography or discussion forums which largely comprise open groups or large group discussions. Therefore, we consider this may, to some extent, limit the degree of interference with rights to privacy in relation to the kinds of services we expect to be in scope of this measure.

15.76 We acknowledge that depending on how age assurance is implemented, for example, whether it is in association with users logging into accounts, having to complete age assurance may result in a user being identified to the service and/or other users via their online account. We recognise that some of the methods of age assurance we have noted may be used for the purposes of this measure, such as those reliant on use of identity documentation, could be more likely to have such impacts. We would however stress that identity verification and age assurance are two distinct concepts, and it is possible to assure a user's age without retaining data other than as needed for the purposes of the age check. We are not recommending that service providers should obtain or retain any specific types of personal data about individual users as part of their highly effective age assurance processes, and in our proposed approach for highly effective age assurance we are giving providers flexibility as to the methods they use, rather than specifically recommending they should rely on identity documentation. We consider that service providers can and should implement the measure in a way which minimises the amount of personal data which may be processed or retained, beyond what is required for implementing the age assurance process, so that it is no more than necessary.

15.77 In processing users' personal data for the purposes of complying with all measures in this section (or in any additional ways they may choose to do so which we are not specifically suggesting[89]), services would need to comply with relevant data protection legislation. This would include abiding by data minimisation principles which require that services collect no more personal data than needed for the purpose of carrying out highly effective age assurance. Data protection legislation also requires they should apply appropriate safeguards to protect the rights of both children and adults. Providers may also use third parties to carry out age assurance; ICO guidance is clear that services should ensure that

[89] For example, if they choose to require adult users to create accounts so that they do not have to repeat age assurance each time they use the service.

individuals' rights to privacy are fully protected when a third party has access to their personal data.[90]

15.78 If a service uses automated processing as part of their highly effective age assurance process (which we are not specifically recommending), we consider that there is a potentially more significant impact on users' rights to privacy, especially if they are unaware that their personal data will be used in this way. Services should refer to ICO guidance to determine whether the processing is solely automated i.e. has no meaningful human involvement, and results in decisions that have a legal or similarly significant effect on users.[91] When implementing age assurance, service providers should have regard to the ICO Commissioner's Opinion on age assurance for the Children's code[92], and comply with the standards set out in the ICO's Age appropriate design code[93] in respect of children's personal data, along with other relevant guidance from the ICO.[94]

15.79 Users' rights in relation to data protection would also be affected by the nature of the action taken as a result of the highly effective age assurance process, particularly if a user's age was incorrectly assessed with the result that personal data held by the service about that user was inaccurate. However, as noted above, while there is potential risk for a margin of error in the use of highly effective age assurance, we propose to recommend that services take account of our recommendations at sub-section 'Our approach to highly effective age assurance' and in Annex 10 (draft HEAA guidance) to ensure the age assurance method implemented is done so in a way that is highly effective. Where incorrect assessments of age are made by a service, the Information Commissioner's Opinion on Age Assurance explains services "must provide tools so that people can challenge inaccurate age assurance decisions. You should make these tools accessible and prominent, so people can exercise their rights easily."[95] Service providers will therefore need to comply with data protection law, following the ICO's Children's code and consulting relevant ICO guidance. This may also mitigate the impact on the adult user's privacy rights or under data protection law by giving the user a mechanism for redress and providing a route to rectify negative impacts by allowing adult users access to the service.

15.80 We therefore consider that the impact of the proposed measure as a result of services' implementation of highly effective age assurance on child and adult users' rights to privacy, to be potentially significant. However, assuming service providers also comply with data protection legislation requirements, it is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the significant benefits to children into consideration, we provisionally conclude that the interference with users' rights to privacy is therefore proportionate.

---

[90] Further information on the requirements for contracts between data controllers and processors can be found at ICO, Contracts and liabilities between controllers and processors. [accessed 18 April 2024].
[91] ICO, Automated decision-making and profiling. [accessed 18 April 2024].
[92] ICO, Children's code guidance and resources for the Commissioner's Opinion on Age Assurance. [accessed 18 April 2024].
[93] ICO, Age appropriate design: a code of practice for online services. [accessed 18 April 2024].
[94] Such as: ICO, Online safety and data protection. [accessed 18 April 2024].
[95] See, for example, the ICO Commissioner's Opinion on Age Assurance, section 6.1.2 [accessed 23 April 2024] and Article 16 UK GDPR.

Measure AA2: Use HEAA to prevent children accessing services whose principal purpose is the hosting or dissemination of PC if the service is also high or medium risk for PC

Services whose principal purpose is the hosting or the dissemination of one or more kinds of PC, and who are high or medium risk for one or more of those kinds of PC, should use highly effective age assurance to prevent children from accessing the entire service.

Explanation of the measure

15.81 Services in scope of Measure AA2 should prevent users from accessing the service unless they have been determined to be adults.

15.82 This measure applies to services whose principal purpose is to host or disseminate one or more kinds of PC, where the service is high or medium risk for one or more of those kinds of PC appearing. As under Measure AA1, 'principal purpose' in this context refers to the main activity or objective of the service.

15.83 It is for the service provider to assess the nature and purpose of its service to determine whether its principal purpose is the hosting or dissemination of PC. Relevant indicators could include:

- Whether the service promotes or refers to PC, for instance through its name, branding, terms of service, or any other means of describing the service to users; and how it markets itself or positions itself against its competitors;
- How the content itself is presented or described, including consideration around whether PC is the main draw for users of the service; or,
- Whether it provides content other than PC.

15.84 In cases where it does, it may be relevant to consider the centrality of PC to the service, including the proportion or relative prominence of PC on the service. We would expect the principal purpose of a service to be the hosting or the dissemination of PC where the content on the service, taken overall, consists entirely or predominantly of PC (discounting any illegal content which should be swiftly removed when identified as required under the illegal content safety duties).

15.85 In making this determination, services may find it helpful to consult our Guidance on Content Harmful to Children in Sections 7.4 - 7.8 provide examples of content which Ofcom would, and would not, consider to be PC.

15.86 Our evidence shows that discussion forums, for instance, can act as spaces where communities of users share content surrounding particular, and sometimes more extreme, topics that can fall in scope of PC.[96] We expect Measure AA2 could include, for example, discussion forums dedicated to gore and violence, or to abusive and hateful content (such as discussion groups set up specifically to degrade or humiliate a target).[97] We would welcome

[96] See Harms guidance section 7.6.
[97] See Harms guidance section on abuse and hate, and bullying in Section 8.7 and 8.10.

further evidence from stakeholders on which types of service host or disseminate PC as their principal purpose and may be in scope of this measure.

15.87 If a proportion of service's residual content is not considered harmful to children (i.e. it is not PPC/PC or NDC), services may wish to consider whether they can create a child-safe experience on their service. We would consider that this is unlikely to be possible if only a minority of residual content is not harmful to children.

15.88 Services have a duty to protect children in age groups judged to be at risk of harm from encountering PC.[98] To secure this duty under Measure AA2, the service provider should implement effective access controls to the entire service to prevent access to users that have not been determined to be adults. To do so, the service provider should implement highly effective age assurance and effective access controls at the point of entry to the service and/or ensure that no regulated user-generated content is visible to users on entering the site before they have completed an age check.

15.89 As under Measure AA1, the effect of this should be that it will no longer be possible for children to normally access the service, and so the service will no longer be likely to be accessed by children. Our Section 4 children's access assessment provides further information on carrying out a new assessment.

Effectiveness at addressing risks to children

15.90 The Act deems PC to be harmful to children and sets out that services have a duty to protect children in age groups judged to be at risk of harm from PC from encountering it.[99]

15.91 The impacts of encountering PC are wide-ranging, extending from emotional harms to loss of life. We discuss the impact of PC harms in more detail in Sections 7.4, 7.5, 7.6, 7.7 and 7.8 in Volume 3 the causes and impacts of harms to children.

15.92 Services whose principal purpose is the dissemination of PC are highly likely to provide unfettered access to this content. We expect children would be almost certain to encounter this content if they accessed these services. We considered whether to recommend that such services use highly effective age assurance to support targeted access controls to specific content or parts of the service, or content control measures to protect children from encountering PC identified on the service (see Measure AA4). However, we considered that these controls would be unlikely to be effective at protecting children from encountering PC on this type of service given that PC will make up at least the majority of content on the service by definition. This approach would require a change in business purpose in ways that are unlikely to be realistic.

15.93 As a result, we consider that preventing access to the entire service using effective service-wide access controls is the only feasible solution to protect children from encountering PC in practice.

15.94 We considered whether to recommend a lower level of effectiveness of age assurance (i.e. lesser than 'highly effective') for this measure to address the different standards of protection outlined in the Act between PPC and PC. We have explained why we do not consider that to be appropriate a in the 'Options Considered' sub-section below. Ultimately, we provisionally conclude that, given the risk of harm that services in scope of Measure AA2

[98] Section 12(3)(b) of the Act.
[99] Section 12(3)(b) of the Act.

present to children, they should have the highest degree of certainty in the age of their users to ensure that children are not misidentified as adults and given unlimited access to explore those services. We consider that outcome would be inconsistent with the objectives of the Act.

15.95 We have also considered whether it would be possible to recommend that services tailor this measure so that access to the service would only be prevented by age groups judged to be at risk of harm, as identified in the service's children's risk assessment. However, we currently have limited evidence linking specific PC harms to different age groups. We will continue to review this and discuss this more in our 'Age Groups' sub-section of this section below.

Rights assessment

15.96 This proposed measure recommends that services whose principal purpose is to host or disseminate PC, and that have a medium or high risk of one or more of those types of PC appearing, should use highly effective age assurance to prevent children from accessing the entire service. In considering the degree of the potential impact on users' and services providers' rights and whether it is proportionate, we have taken as our starting point the requirements of the Act. The children's safety duties set out in the Act require providers of U2U services to use proportionate systems and processes designed to protect children from encountering PC.[100] As set out above, we consider the services in scope of this proposed measure pose a high risk to children encountering PC as we expect the content present on these services to consist entirely or predominantly of PC. As discussed above, evidence shows that the impact of encountering PC could cause serious harm to children's physical, mental or emotional wellbeing. We therefore consider that a substantial public interest exists in measures which aim to protect children from encountering this kind of harmful content. This proposal is designed with a degree of flexibility based on our criteria-based approach to implementing highly effective age assurance which does not mandate a specific method of age assurance.

Freedom of expression and association

15.97 With this proposed measure, potential interference with both child and adult users' rights to freedom of expression and association, and service providers' rights to freedom of expression, arises where the service provider applies highly effective age assurance with the objective of restricting children's access to the entire service to prevent them encountering PC. We consider that the degree of potential interference with these rights is potentially significant for the reasons set out in relation to Measure AA1 above. This would particularly be the case in the event that some services in scope of this measure (for example, smaller services) were to exit the UK market due to the cost burden of implementing age assurance, meaning that UK adults would no longer be able to access these services, or if UK adults are dissuaded by having to complete an age assurance process from accessing them, or are incorrectly assessed to be children and therefore denied access to the service. However, as also outlined above, we have also sought to design the measure to limit these impacts, for example, by proposing that services take account of our recommendations at 'Our approach to highly effective age assurance' and Annex 10 (draft HEAA guidance) to ensure the age

[100] Section 12 (3)(b) of the Act.

assurance method implemented is done so in a way that is highly effective, and is easy to use.

15.98 We also note that this proposed measure may have a significant impact on the freedom of expression and association rights of children who may be in age groups not judged to be at risk of harm from the relevant types of PC. These children would also be prevented from accessing such services, in addition to children who may face much more significant risks if they were able to access them. We recognise that the children's safety duties in the Act place an obligation on services only to protect children in age groups judged to be at risk of harm from encountering PC, and that it is important for us to take into account the different needs of children in different age groups when designing our Codes recommendations, in line with the principles set out in Schedule 4 to the Act. However, for the reasons discussed in our 'Children in different age groups' section below, we are not recommending our measure to be tailored at particular age groups at this time, in particular due to limited evidence on the technical capability for services to place children into age groups below the age of 18 and due to the limited evidence in linking specific PC harms to different age groups. We also note that the severity of impacts faced by children within particular age groups when exposed to PC may vary quite significantly and some children will be more vulnerable than others, even in older age groups such as neurodivergent children and children whose gender, race and sexuality may impact the harm they experience from content outlined in Sections 7.4-7.8 in Volume 3 the causes and impacts of harms to children. Therefore, while there may be some unintended adverse impacts on some children who would be less severely affected if exposed to such content, this may not be the case for all children across a particular age group for which this additional protection may provide significant benefits. As with Measure AA1 above, we also consider that there is a risk that children are prevented from accessing non-harmful content on the service, but given the nature of the services in scope of this measure, we consider that the amount of non-harmful content on the service from which children could potentially benefit from is likely to be very limited.

15.99 Although we recognise the potential for this measure to have a significant impact on users' and service providers' rights to freedom of expression and association, taking into account the nature of services in scope of this measure and the high risk of harm that these services pose to children as set out above, we consider that preventing access to the entire service using effective service-wide access controls is the only feasible solution to provide children with adequate protections from encountering PC on services in scope of this measure in practice. Therefore, we do not consider there is a less intrusive way for services in scope of this proposed measure to meet the requirements of the Act. For all these reasons, and taking into account the significant benefits to children from preventing the harmful consequences of their exposure to PC on these services which may otherwise occur, our provisional view is that the impact on users' and service providers' rights to freedom of expression is therefore proportionate.

Privacy

15.100 We consider that this proposed measure has the potential to have a significant impact on users' (both adults' and children's) rights to privacy and their rights under data protection law for the reasons set out in relation to Measure AA1 above. We consider that the reasons these impacts arise to be the same as set out in Measure AA1 above. In particular, we note that there is a risk this proposed measure could affect children's and adults' ability to access services which provide means for users to communicate privately (e.g. private messaging

functionalities) or communications in relation to which individuals might expect a reasonable degree of privacy. We also note that all methods of age assurance will inevitably involve the processing of personal data of individuals, including children, whose personal data requires special consideration. However, as with Measure AA1, while we have considered carefully whether we should limit this measure such that it does not apply to private communications and/or content communicated privately, we do not consider this to be appropriate because the nature of this services in scope of this proposed measure means that we do not consider that it is likely that children could be prevented from exposure to PC in any private communications functionalities enabled by these services, given the principal purpose of the service would be to host or disseminate content related to one or more kinds of PC. As with Measure AA1, we also consider that any functionalities enabling private communications are likely to be an ancillary function of such services, as we would anticipate that they will generally be focused on open communications – for example, forums dedicated violent content.

15.101 In addition, we have also sought to mitigate the impacts on users' privacy rights through the design of our proposed measure and our proposed approach for the implementation of highly effective age assurance, as set out in connection with Measure AA1 above. In particular, we would reiterate that, in implementing this measure, we expect service providers to have regard to the ICO Commissioner's Opinion on Age Assurance for the Children's code[101], and comply with the standards set out in the ICO's Age Appropriate Design Code in respect of children's personal data, along with other relevant guidance from the ICO.[102] We also expect them to take account of our recommendations at 'Our approach to highly effective age assurance' and Annex 10 (draft HEAA guidance) to ensure the age assurance method implemented is done so in a way that is highly effective and minimises the risks of error.

15.102 In summary, taking into account the nature of services in scope of this measure, and the risk of harm that these services pose to children as set out above, our provisional view is that the potentially significant interference with user's rights to privacy is proportionate to the need to provide an adequate level of protection to children from PC which this measure is designed to secure, in line with the requirements of the Act, provided that service providers comply with data protection legislation requirements.

Impacts on services – Measures AA1 and AA2

15.103 Both proposed Age Assurance Measures AA1 and AA2 recommend that services implement age assurance, whether this is a third-party solution, or a solution built in-house.

15.104 A difference between Measure AA1 and AA2 is that the Act specifically requires services that allow some PPC to be hosted on the service to use highly effective age assurance to prevent children encountering PPC identified on that service (Measure AA1). We therefore consider that the costs to services of implementing age assurance under Measure AA1 result directly from this requirement in the Act.

[101] ICO, Children's code guidance and resources for the Commissioner's Opinion on Age Assurance. [accessed 18 April 2024].
[102] Such as: ICO, Online safety and data protection. [accessed 18 April 2024].

15.105 On the other hand, the Act does not specify how services choosing to allow the hosting of some PC content should satisfy their duty to protect children in age groups judged to be at risk of harm from PC (Measure AA2). Therefore, we are exercising a degree of discretion by recommending the use of highly effective age assurance in relation to identified PC also. Here we set out our impact analysis and cost estimates for implementing highly effective age assurance at a high level, where services are not already required to do so under other safety duties.[103] [104]

15.106 Our cost analysis is based on limited data on current age assurance capabilities and technologies, recognising that these are rapidly evolving and could become more efficient over time. Our analysis reflects that some providers may choose to rely only on third-party age assurance providers, whereas other – likely larger – providers may rely partly or fully on age assurance methods developed internally. For services using a third-party solution, costs are largely expected to scale with the number of users that attempt to access the service, which influences the number of users requiring an age check. However, there are also one-off costs, such as the cost of understanding the highly effective age assurance criteria and preparing for its introduction – which may be less dependent on the size of the service. Therefore, we expect that the total costs of our proposed measures are likely to represent a larger proportion of total revenues for smaller services.

Preparatory costs relating to the introduction of highly effective age assurance

15.107 There are likely to be upfront one-off staff costs relating to understanding the highly effective age assurance criteria and principles that we set out in our guidance, to be able to decide which age assurance method is appropriate.[105] Similarly, the recommendation for providers to familiarise themselves with relevant data protection legislation and ICO guidance pertaining to age assurance may entail some staff costs. As age assurance technology evolves, the service may have to review and update its methods and/or processes to ensure compliance over time.

Direct costs of deployment

15.108 We consider that direct costs are likely to depend on how a service provider approaches implementation of these measures. Where a provider adopts a third-party assurance method, we estimate that most age assurance costs are likely to relate to the ongoing per check costs and therefore depend on how many users the service needs to age check. In contrast, if a service were to build its own age assurance measures, then the cost of age assurance will relate mostly to the significant upfront investment required to build an age assurance process and related ongoing costs.

15.109 In the case of large businesses, it may be more cost effective to develop an age assurance method in house, for example if a provider conducts large volumes of ongoing age checks and if these costs can be spread across multiple services in scope of these measures.[106] We

[103] Services in scope of the Part 5 duties as well as the Part 3 duties would not need to incur additional costs of implementing age assurance but would be able to apply the same systems and processes put in place because of Part 5 duties to meet our requirements set out here under the Part 3 duties.
[104] Based on the labour cost assumptions set out in Annex 12.
[105] These criteria and principles include our guidance on highly effective age assurance which includes principles relating to accessibility, interoperability, and transparency.
[106] If a service can monetise its own in-house age assurance method in other ways may help to make the case for developing an age assurance method in-house also.

consider it less likely that in-house solutions are cost effective for smaller services, who are more likely to make use of third-party age assurance providers.

15.110 Below, Table 15.1 provides illustrative cost estimates for different kinds of services using third-party age assurance providers. As explained in Annex 12, these are based on several stylised assumptions; for example, we assume the cost per age check is constant, whereas in practice large services may be able to secure some level of volume discount.

Table 15.1: illustrative cost estimates of age checks via third-party age assurance providers*[107]

	Existing UK user base	New users each year	Age assurance for existing users (one-off cost)	Age assurance for new users (annual ongoing cost)
Smaller services	100,000	10,000	£5,000 - £20,000	£1,000 - £2,000
	350,000	35,000	£18,000 - £70,000	£2,000 - £7,000
	700,000	35,000	£35,000 - £140,000	£2,000 - £7,000
Larger services	1,000,000	50,000	£50,000 - £200,000	£3,000 - £10,000
	7,000,000	70,000	£350,000 - £1,400,000	£4,000 - £14,000
	20,000,000	200,000	£1,000,000 - £4,000,000	£10,000 - £40,000

Source: Ofcom analysis.

**Note: All cost estimates have been rounded up to the nearest thousand. These stylised examples assume a faster rate of user base growth, in proportionate terms, for the smallest services (10% growth rate) and a lower rate for the largest services (1% growth rate).*

15.111 Where a service adopts a third-party method, we estimate that the one-off cost of checking the age of existing service users could be between £5,000 and £70,000 initially for a service with 100,000 to 350,000 users, between £35,000 and £200,000 for a service with 700,000 to 1 million users, and between £350,000 and £4 million for a service with 7 million to 20 million users.[108] The low estimates are based on an age check cost of £0.05 and the high end estimate on a cost of £0.20 per check, and we assume each user is checked once.

15.112 The annual ongoing cost of checking new users could be between £1,000-£7,000 for a service with 100,000 to 350,000 users, £2,000-£10,000 for a service with 700,000 to 1 million users, and £4,000-£40,000 for a service with 7 million to 20 million users where a service uses a third-party method for these checks.[109] To estimate ongoing costs, we have assumed a faster growth, in proportionate terms, for smaller services than for larger services.[110]

[107] To calculate the cost of age checks, we multiply the number of existing users by the per-check cost (for example, 100,000 existing users x 5p = £5,000).
[108] Detailed assumptions are included in Annex 12.
[109] We assume that ongoing age checks will continue annually as the service adds new users, and that (a) the cost per check remains unchanged over time, (b) all checks for a service cost the same to verify (i.e. any volume discounts are applied to all verified users for that service) and (c) the nature of the service does not influence the per check cost.
[110] While, the total user numbers of a service increase, it does not necessarily mean that the usage grows as some existing users may become less active.

15.113 Where a service chooses to develop an age assurance method internally, the upfront investment is likely to depend on the context of a specific service and its chosen approach. We have developed high-level indicative estimates in the context of a very large business choosing to invest in this, which we consider the more likely scenario. To the extent that smaller services have the relevant capabilities to pursue an in-house approach, it is possible that they may be able to do so in a more cost-effective way than suggested by our indicative cost estimates (e.g. due to having simpler organisational processes and lower overheads in relation to the relevant activities).

15.114 Our indicative analysis suggests that the upfront staff costs relating to development, testing and deployment of an in-house solution could be in the region of many hundreds of thousands and potentially up to £1 million. In addition to these quantified costs, a provider may incur substantial one-off costs relating to acquiring relevant datasets for developing its age assurance method and one-off software/hardware costs relating to additional computational resources to develop and train its age assurance method, which may include cloud infrastructure and data security.[111] We do not quantify these as they are likely to be dependent on the specific age assurance method.

15.115 There would also be ongoing staff costs relating to model monitoring and maintenance. We estimate that these could reach £1 million annually or potentially more, depending on a service's approach. Our estimates are based on the same salary assumptions for upfront and ongoing costs. In practice, it is possible that some ongoing activities could be conducted by more junior staff on lower salaries, such that ongoing costs could be lower than suggested here.[112]

15.116 The measures may also require changes to the service design to control access to the entire service, allowing this only after a successful age check. To develop this will require some software engineers' time. With services choosing a third-party provider method, some age assurance providers include this in their upfront fees. This cost could be in the low thousands of pounds, but this may vary if a service provider's existing systems require a more complex set up for the age assurance method.[113] It is possible that system complexity or other factors could increase these costs, but we still expect these costs to be low compared to direct costs of implementing a system for age checks, which we discuss .

15.117 We note that various testing and evaluation activities are recommended under our highly effective age assurance criteria. Where services use third-party age assurance providers, we have assumed that those third parties would carry out the bulk of these activities, which may limit further costs incurred by services. However, first-party service providers would still be expected to maintain due oversight and understanding of any third-party testing and

[111] While a large service may be able to use existing infrastructure to support development of its new age assurance method, and this way optimise resource utilisation and not incur additional costs, there is an opportunity cost to this which means these resources are not available for other uses.

[112] The upfront staff costs are based on staff input on a full-time equivalent (FTE) basis for six months from 16 employees, while for the ongoing labour costs we assume require 14 FTEs annually. The cost range is based on an annual software engineer pay of £49,430 (low) and £98,860 (high), uplifted by 22% to account for non-wage labour costs, such as employers' National Insurance contributions. This is likely to be an overestimate given that we expect the service to use more junior staff to monitor the model and carry out any maintenance and support functions. Further details are set out in Annex 12.

[113] For example, Yoti charges an initial set up fee of £750 per organisation. Yoti Age Verification Pricing. [accessed 8 April 2024].

evaluation, as it is the service providers in scope of our Age Assurance measures who are ultimately responsible for ensuring that their approach to age assurance is highly effective.

Indirect costs on services

15.118 We recognise that to the extent our proposed measures reduce the number of users on a service, this will adversely affect service providers' revenues. Where revenue impacts are due to excluding children, this is a direct result of the policy intention and for the reasons set out above, we consider this necessary for services whose principal purpose is hosting PPC or PC to meet their duties. However, excluding adults from a service because they do not want or are unable to complete an age check, or are wrongly assessed as children, could also result in lower service revenues over time. We believe that services have the incentive to ensure they choose or develop age assurance technology that facilitates the age check process so that adults engage with it, which should also reduce the indirect cost of it.

Small and micro businesses

15.119 It is possible that some services in scope of these measures may be operated by the same provider, for example where several pornography services share the same parent service provider. We recognise that service providers who operate several services are likely to have an advantage over providers operating a single service, as they may be able to give users access to many services based on a single age check and in this way save on costs or have greater resources to put in place highly effective age assurance compared to single service operators. This is unlikely to apply to small and micro businesses, which could disadvantage them in the short to medium term, including increasing the costs of market entry for new services that cater for users interested in accessing PPC and/or PC.[114] It is possible that decreasing costs of age assurance and greater opportunities for interoperable age assurance systems could mitigate this over time, but this is uncertain. It is possible that because of the cost implications some smaller services may decide to stop serving adult users in the UK.

Which providers we propose should implement these measures

15.120 We consider that, given the extremely high likelihood of children encountering PPC or PC on services whose principal purpose is to host or disseminate such content, our measures have potential to generate a very direct and material positive impact on children's safety online.

15.121 We consider that the risk of severe harm would exist on all services of this nature that focus on PPC and are likely to be accessed by children, and so we provisionally recommend that these measures apply to all services regardless of size or risk. For services whose primary focus is on PC, we consider that the risks for children are likely to be significant in most cases, even if they reach a relatively small number of child users, especially in younger age groups. However, to the extent that low-risk services whose principal purpose is related to PC might exist, then the benefit to children of applying Measure AA2 to those services would be limited. Therefore, we propose that Measure AA2 is subject to the additional condition that services have medium or high risk for at least one kind of PC that their principal purpose is to host or disseminate.

[114] Under the VSP regime, some adult sites closed because of the expectation of having to implement age assurance.

15.122 We are not currently proposing to recommend this measure for services whose principal purpose is to host or disseminate NDC. This is because we have more limited evidence at this stage about services and harm associated with NDC.

15.123 In our assessment we considered that many of the services in scope of our measures are likely to be smaller services, and our proposed measures may be relatively costly for them. While the direct costs are likely to scale with the number of users on a service, we recognise that some services may not be able to carry the direct and indirect cost burden of implementing age assurance and may decide to exit the UK market, which may leave adult users in the UK with less choice to the extent that they benefit from such services. Our measures will also make it more cumbersome for adults to access these services. Services may reduce this 'hassle factor' by requiring all users to register with the service and conduct a one-off age check, as this may also mean costs on services are lower. However, this may reduce the ability for users to access these kinds of services without being logged in. This could have privacy impacts or dissuade adult users from using these services, as noted above. We also acknowledge that the measure would result in all children losing access to these services, including their access to any potentially non-harmful and beneficial content on these services and in respect of any children who would not necessarily be severely harmed by doing so (see further the 'Children in different age groups' sub-section).

15.124 However, as explained in previous sub-sections, we consider these measures to be the only feasible way to secure providers of these kinds of services comply with the duties set out in the Act, with clear potential to substantially improve children's safety online even in respect of smaller services. Overall, we expect it is likely that a number of services in scope of these measures would continue to serve UK adult users. Our criteria-based approach gives service providers flexibility in how to comply, allowing them to future-proof their systems and respond to their user base and technical developments over time in the most cost-effective way for them, which should benefit all regulated services and mitigate any adverse impacts to some degree. We therefore consider the measures proportionate for these kinds of services, taking into account their potential impacts as summarised above.

15.125 Therefore, we propose that:

- Measure AA1 should apply to services whose principal purpose is the hosting or the dissemination of one or more kinds of PPC; and
- Measure AA2 should apply to services whose principal purpose is the hosting or the dissemination of one or more kinds of PC, and that have medium or high risk for at least one of those kinds of PC.

Provisional conclusion

15.126 The use of highly effective age assurance for services who do not prohibit PPC is mandated by the Act, and we have closely reflected that requirement under Measure AA1. Given the harms that Measure AA1 seeks to mitigate in respect of suicide, self-harm and eating disorder content, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes.

15.127 We have exercised a degree of discretion in recommending the use of highly effective age assurance to prevent children from accessing a service in its entirety where the principal purpose of the service is to host or disseminate PC content. Given the harms that Measure AA2 seeks to mitigate in respect of abusive, bullying or violent content, and content which incites hatred or encourages dangerous challenges / substance misuse, and the focus of this

measure on the riskiest services who are high or medium risk for the kinds of PC they host as their principal purpose, we consider this measure to be appropriate and proportionate to recommend for inclusion in the Children's Safety Codes.

15.128 For the draft legal text for these measures, please see PCU H2 and H3 in Annex A7. Please also see Annex 10 (draft HEAA guidance).

Content control measures

15.129 Content controls are mechanisms to determine the visibility and accessibility of content within a service, including its removal or reduction.

15.130 Children have rights to information and participation in the digital world. Many services will be able to create age-appropriate experiences for child users if they are aware which of their users are children and can prevent their exposure to harmful content accordingly. These services should work to establish the age of their child users to prevent access to specific pieces of harmful content or to parts of the service hosting identified harmful content.

15.131 At the same time, services should use content controls to protect children from encountering content that may be harmful to them. Measures AA3 and AA4 recommend the use of highly effective age assurance to facilitate content control measures.

Measure AA3: Use HEAA to prevent children's access to PPC on services that do not prohibit PPC

Services whose principal purpose is not the hosting or the dissemination of one or more kinds of PPC, but which do not prohibit one or more kinds of PPC, should use highly effective age assurance to ensure children are prevented from encountering PPC identified on the service.

Explanation of the measure

15.132 Services who do not prohibit one or more kinds of PPC are required by the Act to use highly effective age assurance to prevent children from encountering PPC identified on the service.[115] Measure AA3 recommends the use of highly effective age assurance to reflect this duty. Users should be prevented from encountering PPC identified on the service if the service has not been able to establish if the user is an adult by means of highly effective age assurance, either because they have established that the user is a child or the user has not completed an age assurance process. This is to prevent all users who either are children or who are not yet determined to be adults from encountering identified PPC. It will also facilitate a more age-appropriate experience for children on the service as they would still have access to other forms of content.

[115] Section 12(3)(a) and 12(4)-(6) of the Act. When identifying these types of content for the purposes of this measure, service providers have a choice: they may either use the categories of content defined in their terms of service, which should be at least as broad as those defined in the Act, or they should use the categories defined in the Act. We provide guidance at Section 8, Volume 3 which gives examples of content, or kinds of content, that Ofcom consider to be (or not to be), primary priority content.

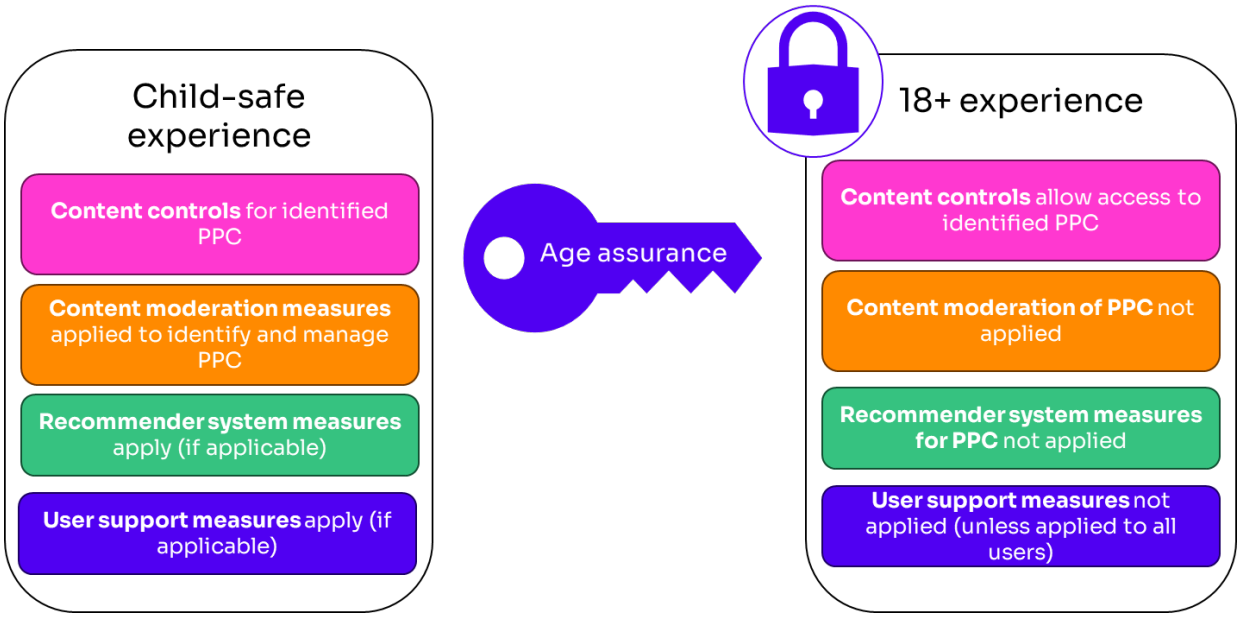
15.133 Measure AA3 will apply to services who do not prohibit one or more kinds of PPC in their terms of service and whose principal purpose is not the hosting or dissemination of PPC (as in such a case, Measure AA1 would apply instead). We expect such services to also host a significant amount of non-harmful content as part of their offering. This may include, for instance, social media services or discussion forums where users upload and share content relating to a wide range of topics, as well as certain kinds of PPC (e.g., pornography).

15.134 One way of achieving the outcome that children are prevented from encountering identified PPC would be to implement age assurance to prevent access to this content. Services could ring fence any PPC which they choose to host on the service, whether that is through applying access controls preventing access to specific pieces of content, or to dissociable parts of the service which host PPC. Services can exercise their discretion in deciding how to implement the content controls so long as the outcome is to prevent users who have not been determined to be adults from encountering identified PPC where it is hosted on a service.

15.135 It is for the service provider to determine where to position the age assurance process on their service, whether at the point of access to the service, account creation, or before a user accesses the part of the service hosting PPC. At whatever point users are required to undergo the age check, under Measure AA3, the service provider should ensure that children (whether logged in or out) are prevented from encountering PPC identified on the service.

15.136 Content moderation will be essential to ensuring that service providers can identify PPC on their service and apply content controls accordingly. For the parts of and content on the service which a child can access, services should ensure children are not able to encounter identified PPC. To do so, services should continue to implement the content moderation measures as per Section 16 (content moderation for U2U services). If a service has a recommender system that children can access, it should also apply Measure AA5 to filter out content likely to be PPC from the spaces where children can access. Users that have not been determined to be adults should have a child-safe experience on the service, as demonstrated in Figure 15.1 below.

Figure 15.1: How Measure AA3 works to create a child-safe experience for users not identified to be adults through age assurance.



Effectiveness at addressing risks to children

15.137 The Act deems PPC to be harmful for children and sets out that services should use proportionate systems and processes designed to prevent children from encountering it. We discuss the impacts on children of encountering PPC in Volume 3, Sections 7.1, 7.2 and 7.3.

15.138 Under Measure AA3, services who do not prohibit one or more kinds of PPC and whose principal purpose is not the hosting or dissemination of PPC, should prevent users from accessing identified PPC unless they have been determined to be adults. The Act mandates that the age assurance used by services who do not prohibit PPC in their terms of service must be highly effective.[116] We do not have discretion to recommend the use of any form of age assurance in these circumstances which is less effective, and so Measure AA3 closely reflects the requirement under the Act.

15.139 We considered recommending the use of age assurance to prevent access to the entire service for all services who did not prohibit PPC. Services whose principal purpose is to host PPC would likely not be able to create child safe environments by using content controls on their service as outlined in Measure AA1. Conversely, services that do not host or disseminate PPC as their principal purpose, and instead host a range of content which children have a right to access and can benefit from alongside allowing one or more kinds of PPC, should be better placed to allow children on their service while preventing them from encountering PPC through content controls.

15.140 In light of this, we are proposing to recommend that where services do not prohibit PPC but do not host or disseminate PPC as their principal purpose, they should use age assurance to enable the use of content controls to prevent children's access to identified PPC, or the parts of the service hosting identified PPC. Service providers can still choose to meet the outcome required by Measure AA3 by preventing access by children to the entire service if they consider it more appropriate.

[116] Section 12(3)a), (4), (5) and (6) of the Act.

15.141 We acknowledge that the effectiveness of Measure AA3 at addressing the risks to children presented by PPC will ultimately depend on how effective a service's content moderation systems and processes are at identifying this type of content. If ineffective systems are used, content may be categorised incorrectly, and children would be exposed to PPC or could lose access to age-appropriate content which the service has wrongly identified as PPC. This may have negative implications on their rights: see sub-section 'Rights Assessment' below.

15.142 We have also recommended that content moderation measures related to the children's safety duties are applied wherever children are on the service which will help ensure that the content a child can access on a service does not include identified PPC. We discuss these measures, and how they effectively address the risks to children of encountering PPC in Section 16. If the content is illegal or in breach of the service's own terms of service, we expect them to remove access to this content for all users.[117]

15.143 If a service chooses to place access controls to parts of its service which host PPC, the effect of this should be that it is no longer possible for children to normally access that part of the service and so it will no longer be likely to be accessed by children. Parts of a service that are not likely to be accessed by children are not in scope of the children's risk assessment and safety duties.[118]

Rights assessment

15.144 This proposed measure recommends that all services that are likely to be accessed by children and do not prohibit PPC are required to use highly effective age assurance to ensure that children are prevented from encountering PPC that the service provider identifies on the service. This measure is designed to have a degree of flexibility and in line with our criteria-based approach it does not mandate a specific method of age assurance.

15.145 By preventing children's access to identified PPC, the proposed measure will seek to secure adequate protections for children from harm, in line with the legitimate aims of the Act. Preventing children from encountering PPC acts to prevent them from experiencing the harmful consequences of such content. These consequences can include harm to children's physical, mental or emotional wellbeing.

Freedom of expression and association

15.146 We consider that this proposed measure has the potential to impact on users' (both adults and children) rights to freedom of expression and of association, and service providers' rights to freedom of expression, for the reasons set out in relation to Measure AA1. Unlike Measure AA1, and for the reasons set out above, this proposed measure does not recommend service providers use age assurance to prevent child users from accessing services altogether. In this respect, the potential impact on children's rights to freedom of expression and association is less significant, as they would still be able to benefit from encountering other (non-harmful) content and interactions on the service, which we consider is in their interests provided they can enjoy an age-appropriate experience on the service while doing so.

---

[117] Services should refer to the draft Online Safety Guidance on Judgement for Illegal Content when determining whether content is illegal. Services may refer to Ofcom's draft Guidance on Content Harmful to Children for examples of content that Ofcom considers to be, or not to be, PPC.

[118] See our proposals for children's access assessments at Volume 2.

15.147 As with Measure AA1, the duty to use highly effective age assurance to prevent children from encountering PPC identified by the service is a requirement of the Act.[119] To the extent that the result of implementing the proposed measure is that children are effectively prevented from encountering PPC identified on the service, and adults are restricted from sharing such content with children, we therefore consider this is the minimum action required to secure that services meet their duties under the Act. However, the proposed measure allows services flexibility as to precisely how age assurance is used to achieve this outcome, and we acknowledge that the precise way that services choose to implement this measure may have a more or less intrusive impact on users' rights to freedom of expression – including of both children and adults.

15.148 We consider that there is a risk, as with Measure AA1, that the significant costs of implementing this measure could mean that some services (for example smaller services) decide to exit the UK market, which would mean that both adults and children in the UK would no longer be able to access these services, although we consider that this is less likely to arise than under Measure AA1 as there is no expectation that this would prevent children's access to the entire service. We also consider there is a risk that some services might choose not to allow any child users in the UK on the service at all, for example if it is difficult or costly for a given service to restrict access only to relevant content or parts of the service.[120] In both cases this would have a potentially significant impact on users' (both children's and adults') rights to freedom of expression and association. However, we have given service providers flexibility as to how to implement this measure in a way which minimises the costs so far as possible, and as noted above, we consider that this measure is closely aligned to the duties in the Act, and we do not consider it likely there is a less intrusive way for services that allow one or more forms of PPC to comply with their duties.

15.149 In relation to children, we acknowledge that one way that service providers might choose to implement the measure would be to restrict their access to dissociable parts of the service where PPC is hosted (i.e. to limit their access to distinct parts of the service). In this case, there could potentially also be restrictions on children encountering other non-harmful content or interacting with other users on those restricted parts of the service. If a service restricted access to non-harmful content for children, this may have additional freedom of expression and association impacts on children. However, we note that this is not something we are specifically recommending as part of this measure - its purpose is to prevent children from encountering identified PPC - and it will be open to services to ensure that they implement it in a way which does not necessarily restrict children's access to non-harmful content. In addition, in the event that services choose to take this approach as part of giving children a more age-appropriate experience on the service overall, this could have significant benefits for them, as it might protect them from other forms of harm as well.

15.150 As outlined in Content Moderation Measure CM1 in Section 16, we also recognise that children's rights to access non-PPC might be impeded in the event that content that is not PPC is wrongly identified as such and, as a result of this proposed measure, they are unable to encounter it. However, as also explained Content Moderation Measure CM1, services have incentives to limit the amount of content that is wrongly actioned, to meet their users'

[119] Sections 12(3)(a), 12(4) and 12(6) of the Act.

[120] We note, however, that preventing children from accessing the whole service may involve additional age assurance costs, as age checks would then be needed for all users who wish to access the service, as well as potentially reducing the total number of users on the service, which may discourage many services from taking this approach.

expectations and to avoid the costs of dealing with appeals. In addition, the complaints procedures outlined in Section 18 should allow for the user to complain and for appropriate action to be taken in response, and this may also give a mechanism for redress. In addition, where services are in scope of Content Moderation Measures CM2-7 in Section 16 and adopt these measures, they should also help to limit the risks that content is wrongly classified as PPC and children's access to it is wrongly restricted as a result.[121] In respect of recommender systems, we discuss in relation to Measure AA5 below the likely impacts connected to that measure, and we do not address them separately here.

15.151 In relation to adults, our measures will make it more cumbersome for adults to access the PPC allowed on these services. As discussed in Measure AA1, some services may make their service available only to users with accounts, to reduce costs by requiring a one-off age check, which may result in a reduced ability for adults to access this content without being logged in. However, they will still be able to encounter all content on the service if they do create an account and complete the age assurance process. Equally, services may still offer users a logged-out experience, however, as per this measure, services would have to ensure this experience does not host identified PPC. Furthermore, as per Measure AA4, they would have to protect all users from identified PC in order to protect potential logged out child users. We also consider users are, on balance, less likely to be dissuaded from accessing the service at all compared with Measures AA1 and AA2, as they may end up losing access to a wide range of beneficial content and user interactions if they choose to forgo this, not just PPC. Where services choose to implement the age assurance process so that adult users are not required to undergo age assurance unless they wish to make a specific choice to access PPC, we acknowledge it is also possible that some adult users might prefer not to complete age assurance and therefore may be dissuaded from seeking to access PPC on the service as a result. We consider this impact on their freedom of expression and association rights to be relatively limited, given they will have a viable option to access the content if they assure their age, and it would therefore be their choice not to follow this mechanism. As noted above, we consider these risks will also be potentially limited by the fact that service providers have incentives to make their age assurance process as user-friendly as possible and limit friction to adult users. This would also limit the risk that some adults may find it more difficult to assure their age under certain methods. We have also reflected the importance of this via our proposed approach to implementing highly effective age assurance, in that we propose services should consider the principle that age assurance should be easy to use including by children of different ages and with different needs.[122]

15.152 As with Measure AA1, we note that adult users' access to PPC might also be inadvertently restricted if they are incorrectly assessed to be children. While there is potential risk for a margin of error in the use of highly effective age assurance, we consider this risk to be limited provided that services take account of our recommendations at 'Our approach to highly effective age assurance' and Annex 10 (draft HEAA guidance) to ensure the age assurance method implemented is done so in a way that is highly effective. Where incorrect assessments of age are made by a service, complaints procedures required under section 21(2) of the Act (as outlined in Section 18) should allow for the user to complain, and for

[121] Unlike Measures AA1 and AA2, services for which Measures AA3 or AA4 are recommended are potentially likely to be accessed by children so will be within scope of this code.

[122] Age assurance should be easy to use and work for all users, regardless of their characteristics or whether they are members of a certain group. Please refer to the HEAA Annex for the practical steps for services to consider.

appropriate action to be taken in response. The complaints process may also mitigate the impact on the adult user's rights to freedom of expression and freedom of association by giving the user a mechanism for redress and providing a route to rectify negative impacts by allowing adult users access to the service.[123]

15.153 We also note that this proposed measure may have unintended impacts on adult users' rights to access to PPC in the event that, to avoid the costs associated with this measure, services choose to change their terms of service to prohibit all forms of PPC. This could have impacts on their rights to freedom of expression and association. However, it remains open to services as a commercial matter (and in the exercise of their own right to freedom of expression) to decide what forms of content to allow or not to allow on their service so long as they comply with the Act. We consider the specific implications that this may have in connection with private communications below.

15.154 The proposed measure could also have positive impacts on freedom of expression and freedom of association rights of children, for example, it could result in safer spaces online where children may feel more able to join online communities and receive and impart (non-harmful) ideas and information with other users. This measure could therefore also have significant benefits to children, in terms of safeguarding their rights to freedom of expression and association in safer online spaces, as well as in terms of protecting them from exposure to harm.

15.155 While we recognise the potentially significant impacts on freedom of expression and association outlined above, as also explained above, the proposed measure is likely to constitute the minimum degree of interference required to secure that service providers who do not prohibit PPC fulfil their children's safety duties under the Act. Taking this, and the significant benefits to children into consideration, we consider that the interference with users' and service providers' rights to freedom of expression and association is therefore proportionate.

Privacy

15.156 We consider that this proposed measure has the potential to impact on users' (both adults' and children's rights) to privacy for the reasons set out in relation to Measure AA1 above.

15.157 As set out in Measures AA1 and AA2 above, all age assurance processes will inevitably involve the processing of personal data of individuals, including children. It will therefore impact on users' rights to privacy and their rights under data protection law. The degree of interference will depend to a degree on the extent to which the nature of their affected content and communications is public or private, or, in other words, gives rise to a legitimate expectation of privacy. It will also depend on the nature of the information required to complete the highly effective age assurance process, for example, the more sensitive information required the more intrusive the method of highly effective age assurance is likely to be.

15.158 As with Measures AA1 and AA2 above, this proposed measure is not limited only to content or communications that are communicated publicly, and may lead to impacts on children's – and for the reasons noted above, adults' – ability to access services, parts of services, or content or communications in relation to which individuals might expect a reasonable degree of privacy. This would in turn lead to more significant privacy impacts than in

[123] See Measure UR1 in Section 18.

connection with impacts on content and communications that are widely publicly available (whether on the service concerned or more generally). The impact on users' rights would also be affected by the nature of the action taken to implement this proposed measure. For example, the level of intrusion and significance of the impact is likely to be higher where services enabling private communications withdraw from the UK, where services prohibit the sharing of PPC on the service, where adults choose to avoid using services enabling private communications due to requirements to complete age assurance, where children are restricted from communicating privately in particular group chats or closed user groups due to suspected risks they are being used to share PPC, or where specific restrictions are put in place on the sharing of PPC via private communications (e.g. forwarding to child users and/or removing from identified content that has already been shared from private communications so that it cannot be further shared). To the extent that these restrictions would potentially impact on children's ability to communicate with their family members, this could also affect their rights to a family life.

15.159 Some of these impacts on rights to privacy and a family life would only follow as a result of essentially commercial choices made by services (as noted in the discussion above), although this would not make them more limited in their impact. Others, however, would be likely unavoidable if services were to implement the measure effectively – for example, preventing children's access to PPC via use of age assurance by limiting their ability to interact in particular private groups or with particular content communicated privately. We have considered carefully whether we should limit this measure such that it does not apply to private communications/content communicated privately to mitigate these potentially significant impacts. We have decided not to do so for two reasons. Firstly, the Act is clear that these services and content are in scope of the children's safety duties, and also requires that services who do not prohibit one or more forms of PPC must use highly effective age assurance as part of their systems and processes to prevent children from accessing PPC they allow. Second, Volume 3, Sections 7.1-7.3, Children's Register of Risks, highlights group messaging as a key functionality through which harmful content is shared among children, and in some cases this risk might arise in connection with group chats in relation to which users would expect a legitimate expectation of privacy. We therefore consider that to the extent this measure can apply to PPC identified in private communications in a proportionate way, this would be consistent with the Act and the risk to children that group messaging functionalities might pose.

15.160 We recognise that children as well as adults have a right to private communications and services should design the application of this measure in a way that limits the impacts on privacy to no more than necessary to give the effect to the requirement of the Act. Within these services, new methods are being developed and deployed to create safer environments including in private communications. As per our fifth consultation question on this section, we welcome responses on the scope of these measures and how they might be implemented in private communications.

15.161 As noted above in relation to Measure AA1, there are also particular risks in relation to privacy and personal data if the age assurance methods deployed by service providers result in the processing of more personal data than needed, or if users' ages are incorrectly assessed, including in relation to content communicated publicly, for example adult users being prevented from encountering this content. This could result in services (and third-party providers of highly effective age assurance) holding unnecessary amounts of users' personal data or having inaccurate personal data of users. As set out in Measures AA1 and

AA2 above, we consider this risk can be mitigated by services having in place appropriate complaints policies and processes as set out in Section 18 (user reporting and complaints). Services will also need to comply with data protection laws and ICO guidance to ensure that users are able to fully exercise their rights in respect of their personal data as we have set out above in Measure AA1.

15.162 We therefore consider that the impact of the proposed measure because of services' implementation of highly effective age assurance on child and adult users' rights to privacy, to be potentially significant. However, assuming service providers also comply with data protection legislation requirements, it is likely to constitute the minimum degree of interference required to secure that service providers that do not prohibit PPC fulfil their children's safety duties under the Act. Taking this, and the significant benefits to children into consideration, we consider that the interference with users' rights to privacy is therefore proportionate.

Measure AA4: Use HEAA to protect children from PC on services that do not prohibit PC

Services whose principal purpose is not the hosting or the dissemination of one or more kinds of PC; and which do not prohibit one or more kinds of PC; and are high or medium risk for one or more kinds of PC that they do not prohibit, should use highly effective age assurance to ensure that children are protected from encountering PC identified on the service.

Explanation of the measure

15.163 Services must protect children in age groups judged to be at risk of harm from PC from encountering it on their service. To meet this outcome, Measure AA4 recommends the use of highly effective age assurance to protect users from content identified as PC unless they have been determined to be adults. This means that if a user is not age assured, they should also be protected from encountering identified PC.

15.164 We propose to apply Measure AA4 to services who do not prohibit one or more kinds of PC in their terms of service and whose principal purpose is not the hosting or dissemination of PC (as in that case, Measure AA2 would apply instead), where they are high or medium risk for one or more kinds of PC they do not prohibit from appearing on the service.[124] This may include, for instance, social media services or discussion forums where users upload and share content relating to a wide range of topics, as well as certain kinds of PC allowed on the service. Some services in scope of Measure AA4 may also be in scope of Measure AA3 if they also do not prohibit one or more kinds of PPC.

15.165 To achieve the outcome of protecting children from PC, services should ensure that all child users benefit from the appropriate action they choose to take in relation to identified PC, as described in Content Moderation Measure CM1 in Section 16. In particular, we set out there

[124] When identifying these types of content for the purposes of this measure, service providers have a choice: they may either use the categories of content defined in their terms of service, which should be at least as broad as those defined in the Act, or they should use the categories defined in the Act. We provide guidance at Volume 3, Section 8 which gives examples of content, or kinds of content, that Ofcom consider to be (or not to be), priority content.

a non-exhaustive list of content moderation actions including limiting the prominence of PC. Services who have recommender systems and are high or medium risk of PC should also apply the relevant recommender system measures to significantly limit the prominence of PC (see Measure AA6).[125]

15.166 One of the ways a service may protect children from PC that is listed in the Content Moderation section is by ringfencing identified PC on the service, whether that is individual pieces of content, or dissociable parts of the service which host the identified PC. Only users that have been determined to be adults would be able to access the identified PC, or the parts of the service where it is hosted. In this context, parts of a service could include but are not limited to tabs, forums, communities, groups. Content moderation will be essential to ensuring that service providers can identify PC on their service and apply access controls accordingly. The service provider should apply the content moderation measures (see Section 16) to protect children from encountering PC on any parts of the service they can access (whether logged in or out). Where a service in scope of this proposed measure has a recommender system Measure AA6 would also apply.

15.167 It is for the service provider to determine which additional measures are necessary alongside highly effective age assurance to achieve the outcome of this measure that children are protected from PC.

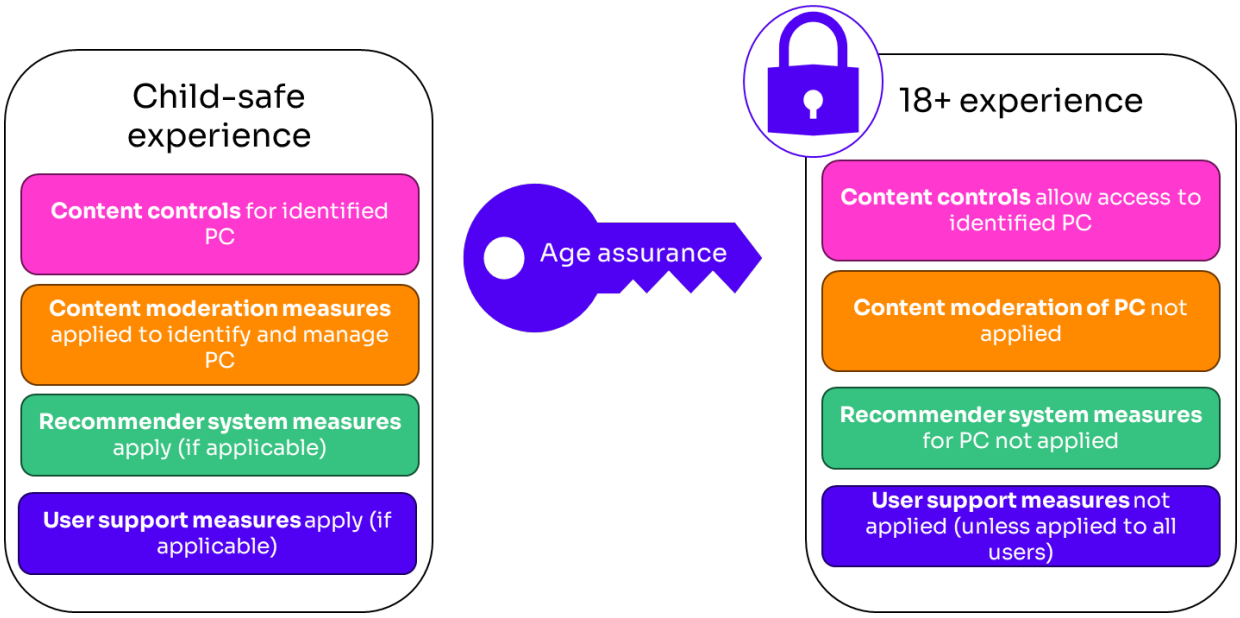
15.168 We considered whether to recommend the use of targeted access controls alone to secure the outcome that children are protected from encountering PC under Measure AA4. Given that the duty on services is to "protect," rather than "prevent" children from encountering PC, we determined it would be more proportionate to give services the flexibility to determine the appropriate action in response to PC, as set out in Content Moderation Measure CM1 in Section 16. The action a service takes may depend on a number of factors, including the nature and severity of the harm.

15.169 Equally, it is for the service provider to determine how to implement a highly effective age assurance process on its service to secure the outcome that children are protected from encountering identified PC. This could be done through age assuring users at the point of access to the service, at account creation or when users attempt to access specific content or parts of the service, for example. Regardless of the position of the age check, the outcome should be that children are protected from encountering PC identified on the service.

15.170 In effect, users that have not been determined to be adults should be protected from identified PC. We demonstrate how this could work using the example of content controls in Figure 15.2 below, although in principle the same approach could be used in respect of other appropriate actions taken: for example, if a service chooses to limit the visibility and prominence of content, e.g. by downranking identified PC, then unless this action is applied to all users, they should use highly effective age assurance to ensure that this is the outcome for users not determined to be adults via their age assurance process.

[125] The Content Moderation Section 1 sets out appropriate actions services can take to protect children from PC and NDC.

Figure 15.2: Example of how Measure AA4 can work to create a child-safe experience for users not identified to be adults through age assurance.



Effectiveness at addressing risks to children

15.171 As discussed under Measure AA2, the Act deems PC to be harmful to children and sets out that services have a duty to use proportionate systems and processes designed to protect children in age groups judged to be at risk of harm from PC from encountering it.[126]

15.172 Services which do not prohibit users from uploading or sharing PC and are high or medium risk for the kinds of PC they do not prohibit on the service are likely to provide access to this harmful content to children if they cannot accurately determine whether a user is a child. Some of these services can host large volumes of PC as part of their offering to users.

15.173 The impacts of PC can be detrimental to children and their wellbeing. These impacts are fully documented in Volume 3 'the causes and impacts of harms to children.' Evidence in Section 7.6 shows that an increase in volume of violent content can lead to children being desensitised to it. Equally where a child encounters a greater volume of PC, this may have a cumulative effect on their wellbeing as also outlined in Section 7.6. PC can include abusive and violent content. Violent content can include violence against women and girls which does not meet the threshold of illegality. PC may also include dangerous stunts and challenges which can pose threats to children's physical safety. While services may host other content that children may benefit from engaging with, this should not come at the cost of having to encounter PC. Services should control the visibility of this content to children to protect them from the harms associated with it.

15.174 Measure AA4 seeks to address this by recommending that services use highly effective age assurance to protect children from encountering this content when identified on the service. The use of highly effective age assurance will provide services with a high degree of certainty

[126] Section 12(3)(b) of the Act.

as to whether or not particular users are children. Services can use this to ensure that only users who have been determined to be adults have uninhibited access to identified PC.

15.175 To apply this measure, services may choose to use content controls[127] to ringfence specific content or parts of the service hosting identified PC, and/or choose to take any other appropriate action, such as limiting the prominence of content to child users where relevant as per Section 16 (content moderation for U2U services).

15.176 If a service chooses to limit access to parts of its service which host PC for users not determined to be adults by means of highly effective age assurance processes, the effect of this should be that it is no longer possible for children to normally access that part of the service. Parts of a service that are not likely to be accessed by children are not in scope of the children's risk assessment and safety duties.[128] As a result, the content moderation measures relevant to children's safety duties would not apply in these spaces.

15.177 We acknowledge that the effectiveness of Measure AA4 at addressing the risks to children presented by identified PC will ultimately depend on how effective a service's systems and processes are at identifying this type of content. If ineffective systems are used, content may be categorised incorrectly and children would be exposed to PC, thereby exposing them to harmful content. Alternatively, children might lose access to or visibility of age-appropriate content which the service has wrongly identified as PC, which may have negative implications on their rights (as discussed under the 'Rights assessment' sub-section below).

15.178 Content moderation measures designed to prevent children from encountering PPC and protect children from PC should apply anywhere a child is likely to access the service. We discuss our approach to these proposed measures, and how they effectively address the risks to children of encountering PC in Section 16 (content moderation for U2U services). If the content is illegal or in breach of the service's own terms of service, we expect them to remove access to this content for all users.[129]

15.179 We considered recommending service-wide access controls to prevent children from accessing services entirely where they do not prohibit PC and are at high or medium risk of it appearing. As outlined in Measure AA2, services whose principal purpose is to host PC would likely not be able to create child safe environments by using content controls on their service. Services that do not host or disseminate PC as their principal purpose, and instead host a range of other content which children have a right to access and can benefit from, are likely to be better placed to allow children on their service by putting in place content controls to protect them from encountering PC. In light of this, we considered that it would be proportionate for such services to use age assurance to facilitate the use of more targeted access controls to protect children from encountering PC. Service providers may still choose to meet the outcome required by Measure AA4 by preventing access by children to the entire service if they consider it appropriate.

---

[127] Content controls mechanisms determine the visibility and accessibility of content including its removal or reduction. In this context, content controls include access controls such as blocking access to a part of the service that may host the harmful content.

[128] As discussed above (measure AA1), where a service implements highly effective age assurance, it can carry out a new children's access assessment to determine whether that part(s) is out of scope of the children's safety duties. See Volume 2, Section 4 and our draft Childrens Access Assessment Guidance at Annex 5 for more information.

[129] Services should refer to the draft Illegal Content Judgements Guidance when determining whether content is illegal.

15.180 We also considered alternatives to highly effective age assurance for determining the age of users, to reflect the different standards of protection outlined in the Act between PPC and PC. We have discussed the problems we encountered in defining a lower level of effectiveness for age assurance at sub-section 'Options considered' below. We provisionally concluded that recommending a lower level of assurance was an inappropriate route for reflecting the different standards of protection. Instead, we have reflected this difference in the Codes in other ways, for instance through affording services greater flexibility in the actions they can take when responding to PC as discussed above.

15.181 We also considered whether it would be possible to recommend that services tailor this measure so that access to the service would only be prevented by age groups judged to be at risk of harm, as identified in the service's children's risk assessment. However, we currently have limited evidence linking specific PC harms to different age groups. We will continue to review this and discuss this more in our 'Children in different age groups' sub-section below.

Rights assessment

15.182 This proposed measure is recommended for services that do not prohibit one or more kind of PC and are high or medium risk of one or more of the kinds of PC that is not prohibited on their service. Protecting children from encountering PC acts to prevent them from experiencing the harmful consequences of such content. These consequences can include harm to children's physical, mental or emotional wellbeing. There is a substantial public interest in this outcome. This proposal is designed with a degree of flexibility based on our criteria-based approach to implementing highly effective age assurance which does not mandate a specific method of age assurance.

Freedom of expression and association

15.183 We consider that this proposed measure has the potential to impact on users' (both adults' and children's) rights to freedom of expression and of association, and service providers' rights to freedom of expression, for the reasons set out in relation to the measures discussed above. Unlike Measure AA2, and for the reasons set out above, this proposed measure does not require service providers to use age assurance to prevent child users from accessing services altogether; services also have more flexibility as to the appropriate action they take in connection with identified PC in relation to this measure compared to Measure AA3, in line with Content Moderation Measure CM1 in Section 16 (content moderation for U2U services). In this respect, the potential impact on children's rights to freedom of expression and association is much significant, as they would still be able to benefit from encountering other (non-harmful) content and interactions on the service, and may still be able to encounter some PC on the service in way that is more proportionate to the risks of harm to them, which we consider is in their interests provided they can enjoy an age-appropriate experience on the service while doing so.

15.184 The duty to use proportionate systems and processes to effectively protect children from encountering PC is a requirement of the Act. To the extent that the result of implementing the proposed measure is that children are effectively protected from encountering PC identified on the service, and adults are restricted from sharing such content with children, we therefore consider this is justified and proportionate to secure that services meet their duties under the Act. However, we acknowledge that the Act does not require the use of highly effective age assurance in achieving this outcome (unlike for PPC), and the proposed

measure allows services flexibility as to precisely how age assurance is used to achieve this outcome. We also acknowledge that the precise way that services choose to implement this measure may have a more or less intrusive impact on users' rights to freedom of expression – including of both children and adults.

15.185 We consider that there is a risk, as with the above measures, that the significant costs of implementing this measure could mean that some services (for example smaller services) decide to exit the UK market, which would mean that both adults and children in the UK would no longer be able to access these services, although we consider that this is less likely to arise than under Measure AA2 as there is no expectation that this would prevent children's access to the entire service. We also consider there is a risk that some services might choose not to allow any child users in the UK on the service at all to avoid the costs of this measure. In both cases this would have a potentially significant impact on users' (both children's and adults') rights to freedom of expression and association. However, we have given service providers flexibility as to how to implement this measure in a way which minimises the costs so far as possible. We also consider it is unlikely to be possible for services in scope of this measure to secure adequate protections for children from PC unless they know who their child users are via the use of age assurance.

15.186 In relation to children, we acknowledge that one way that service providers might choose to implement the measure would be to restrict their access to dissociable parts of the service where PC is hosted (i.e. to limit their access to distinct parts of the service). In this case, there could potentially also be restrictions on children encountering other non-harmful content or interacting with other users on those restricted parts of the service. However, we note that this is not something we are specifically recommending as part of this measure and it will be open to services to ensure that they implement it in a way which has the least possible impact on restricting children's access to beneficial content. In addition, in the event that services choose to take this approach as part of giving children a more age-appropriate experience on the service overall, that could have significant benefits for them, as it might protect them from other forms of harm as well.

15.187 We also note that this proposed measure may have a significant impact on the freedom of expression and association rights of children who may be in age groups not judged to be at risk of harm from the relevant types of PC. These children would also face restrictions on their access to PC on these services, in addition to children who may face much more significant risks from encountering PC. As explained under Measure AA2 above, for the reasons discussed in the 'Children in different age groups' sub-section below, we are not recommending our measure to be tailored at particular age groups at this time, in particular due to limited evidence on the technical capability for services to place children into age groups below the age of 18 and due to the limited evidence in linking specific PC harms to different age groups. We also note that the severity of impacts faced by children within particular age groups when exposed to PC may vary quite significantly and some children will be more vulnerable than others, even in older age groups such as neurodivergent children and children whose gender, race and sexuality may impact the harm they experience from content as illustrated in sub-section 'User Demographics' of Section 7.6 Violent Content. Therefore, while there may be some unintended adverse impacts on some children who would be less severely affected if exposed to such content, this may not be the case for all children across a particular age group for which this additional protection may provide significant benefits.

15.188 As outlined in Content Moderation Measure CM1 in Section 16, we also recognise that children's rights to access non-harmful content might be impeded in the event that content that is not PC is wrongly identified as such and, as a result of this proposed measure, they are unable to, or less likely to, encounter it. The relevant Content Moderation Measures CM1 to Measure AA2 recommends swift action on PPC and PC but give services more flexibility as to how to appropriately action PC (and NDC) on relevant parts of the service. We therefore consider this impact to be more limited than the equivalent impact under Measure AA3. In addition, as also explained in Content Moderation Measure CM1 in Section 16, services have incentives to limit the amount of content that is wrongly actioned, to meet their users' expectations and to avoid the costs of dealing with appeals. Furthermore, the complaints procedures outlined in Section 18 should allow for the user to complain and for appropriate action to be taken in response, and this may also give a mechanism for redress. In addition, where services are in scope of Content Moderation Measures CM2-7 in Section 16 and adopt these measures, they should also help to limit the risks that content is wrongly classified as PC and children's access to it is wrongly restricted as a result. In respect of recommender systems, we discuss in relation to Measure AA6 below the likely impacts connected to that measure, and we do not address them separately here.

15.189 In relation to adults, our measures will make it more cumbersome for adults to access PC that is allowed on these services. As discussed in the above measures, some services may make their service available only to users with accounts, to reduce costs by requiring a one-off check, which may result in a reduced ability for adults to access this content without being logged in. However, they will still be able to encounter all content on the service if they do create an account and complete the age check. Where services choose to implement the age assurance process so that adult users are not required to undergo age assurance unless they wish to make a specific choice to access PC, we acknowledge it is also possible that some adult users might prefer not to complete age assurance and therefore may be dissuaded from seeking to access PC on the service as a result. We consider adult users are, on balance, less likely to be dissuaded from accessing the service at all than under Measures AA1 and AA2, as they may end up losing access to a whole range of beneficial content and user interactions if they choose to forgo this, not just PC. We consider this impact on their freedom of expression and association rights to be relatively limited, given they will have a viable option to access the content if they assure their age, and it would therefore be their choice not to follow this mechanism. In both cases, we consider these risks will also be potentially limited by the fact that providers have incentives to make their age assurance process as user-friendly as possible and limit friction to adult users. This would also limit the risk that some adults may find it more difficult to assure their age under certain methods. We have also reflected the importance of this via our proposed approach to implementing highly effective age assurance, in that we propose to recommend that providers should take account of the principle that age assurance should be easy to use including by children of different ages and with different needs.[130]

15.190 As with Measure AA2, we note that adult users' access to PC might also be inadvertently restricted if they are incorrectly assessed to be children. We consider this risk to be limited provided services implement our recommended principles for highly effective age

[130] Age assurance should be easy to use and work for all users, regardless of their characteristics or whether they are members of a certain group. Please refer to the HEAA Annex for the practical steps for services to consider.

assurance, and also the complaints process that services will be required to make available as per Section 18 (User reporting and complaints).

15.191 We also note that this proposed measure may have unintended impacts on adult users' rights to access to PC in the event that, to avoid the costs associated with this measure, services choose to change their terms of service to prohibit all kinds of PC. This could have impacts on their rights to freedom of expression and association. However, it remains open to services as a commercial matter (and in the exercise of their own right to freedom of expression) to decide what forms of content to allow or not to allow on their service so long as they comply with the Act. We consider the specific implications that this may have in connection with private communications below.

15.192 The proposed measure could also have positive impacts on freedom of expression and freedom of association rights of children, for example, it could result in safer spaces online where children may feel more able to join online communities and receive and impart (non-harmful) ideas and information with other users. This measure could therefore also have significant benefits to children, in terms of safeguarding their rights to freedom of expression and association in safer online spaces, as well as in terms of protecting them from exposure to harm.

15.193 While we recognise the potentially significant impacts on freedom of expression and association outlined above, as also explained above, we consider it to be unlikely that there is a less intrusive way to secure that these services, which would be medium to high risk of one or more kinds of PC that they allow on their service, comply with their children's safety duties under the Act relating to PC. Taking this, and the significant benefits to children into consideration, we consider that the interference with users' and service providers' rights to freedom of expression and association is therefore proportionate.

Privacy

15.194 We consider that this proposed measure has the potential to impact on users' (both adults' and children's rights) to privacy for the reasons set out in relation to the measures set out above.

15.195 As with Measure AA3, we consider that the degree of interference will depend to a degree on the extent to which the nature of their affected content and communications is public or private, or, in other words, give rise to a legitimate expectation of privacy and on the nature of the action taken to implement this proposed measure. We consider that the level of intrusion and significance of the impact to these rights to be similar to those highlighted under Measure AA3. For the reasons set out in Age Assurance Measure 3 (AA3), we consider the potential degree of interference with users' privacy rights to be significant particularly where this may lead to impacts on children's and adults' ability to access services, parts of services, or content or communications in relation to which individuals might expect a reasonable degree of privacy, or if adult users' ages are incorrectly assessed. We acknowledge that there is a risk that adults choose to avoid using services enabling private communications due to requirements to complete age assurance or services choose to restrict children a communicating privately in particular group chats or closed user groups due to suspected risks they are being used to share PC. In these circumstances the impact on rights to privacy may be higher, and to the extent that these restrictions would potentially impact on children's ability to communicate with their family members, this could also affect their rights to a family life.

15.196 The level of interference with this right will also depend on the nature of the information required to complete the highly effective age assurance process, for example, the more sensitive information required the more intrusive the method of highly effective age assurance is likely to be. As mentioned in previous sub-sections, services must abide by existing data protection legislation. When implementing age assurance, service providers should have regard to the ICO Commissioner's Opinion on Age Assurance for the Children's code[131], and comply with the standards set out in the ICO's Age Appropriate Design Code[132] in respect of children's personal data, along with other relevant guidance from the ICO.[133]

15.197 As with the measures above, we have considered carefully whether we should limit this measure such that it does not apply to private communications/content communicated privately so as to limit these potentially significant impacts. We do not consider this to be appropriate given that services in scope of this proposed measure would pose a high to medium risk to PC being present and evidence which shows that group messaging is a key functionality though which PC (for example violent content) is shared amongst children. Therefore, we therefore consider that to the extent this measure can apply to PC identified in private communications in a proportionate way, this would be consistent with the risk to children that group messaging functionalities might pose on the services we proposed to be in scope of this measure. In addition, as discussed above in relation to the above measures, we have sought to mitigate the impact on users' privacy rights through the design of our proposed measure and our proposed approach for the implementation of highly effective age assurance. We also consider that services should design the application of this measure in a way that limits the impacts on privacy to no more than necessary to give the effect to the requirement of the Act, namely securing adequate protections for children from encountering PC. As per our consultation question 5 in our 'Consultation questions' on this section we welcome responses on the scope of these measures and how they might be implemented in private communications.

15.198 Overall, we consider it to be unlikely that there is a less intrusive way to secure that these services, which would be medium to high risk of one or more kinds of PC that they allow on their service, comply with their children's safety duties under the Act relating to PC, provided they also comply with data protection legislation requirements. Taking this, and the significant benefits to children into consideration, we consider that the interference with users' rights to privacy is therefore proportionate.

Impacts on services – Measures AA3 and AA4

15.199 Our proposed Measures AA3 and AA4 would require service providers to implement highly effective age assurance at least for users who seek access to PPC and/or PC. The Act specifically requires the use of highly effective age assurance to prevent children from encountering identified PPC. We therefore consider that the costs to services of implementing age assurance under Measure AA3 result directly from this requirement in the Act. For Measure AA4, we have exercised discretion in recommending highly effective age assurance to protect children from encountering identified PC. Our analysis in the remainder of this section therefore focuses on the cost implications of Measure AA4.

---

[131] ICO, 2024. Age Assurance for the Children's code. [accessed 19 April 2024].
[132] ICO. Age appropriate design: a code of practice for online services. [accessed 19 April 2024].
[133] Such as: ICO. Online safety and data protection. [accessed 19 April 2024].

15.200 We expect the cost of implementing highly effective age assurance to be similar to those discussed in more detail under Measures AA1 and AA2 above. Costs of age assurance can be substantial and are likely to depend on the approach a service takes to implementation (e.g., the method(s) used), and the size and number of users service has. However, whereas Measures AA1 and AA2 would mean implementing highly effective age assurance for every user accessing the service, Measures AA3 and AA4 may only require age checks for a subset of users. This could limit the costs, depending on the number of users who choose to undertake an age check to access identified PPC or PC, which will vary depending on the nature of the service and its user base. Some service providers may choose to age check all users at a sign in stage or registration, which would increase costs, but this is at service's discretion and the proposed measure does not specifically recommend such an approach.

15.201 If a service implements age checks when users seek access to PPC/PC, costs will generally be higher if many users are motivated to undergo an age check. We consider that this is more likely to be the case on services with large volumes of PPC/PC, whereas users may be less motivated if the age check only unlocks limited volumes of additional content. Such services with large volumes of PC would tend to be riskier, all else equal. Therefore, we consider that, on average, costs may tend to be higher for riskier services, where the measure also has greater potential to benefit children through increased safety.

15.202 Service providers may also incur costs to take appropriate action with respect to protecting children from identified PC (for example, restricting access to dissociable parts of the service where identified PC is prevalent, or applying content blurring or filtering to specific pieces of content for child users). Such costs would depend on the context of each specific service and the actions taken. They would primarily result from the separate Content Moderation Measure CM1 in Section 16 which involves taking appropriate action, whereas the costs of the proposed Measure AA4 primarily concern the approach to identifying user age for the purpose of targeting appropriate actions at children. Any costs associated with reducing the prominence of PC on recommender systems would result from the separate Recommender Systems Measure RS2 in Section 20 and Measure AA6 rather than the proposed Measure AA4.

15.203 Services may also experience indirect costs through reduced user engagement, user numbers and revenue if there are adults who are discouraged from using the service because they want to access identified PPC and PC, but they are unwilling to complete the age assurance process. As discussed for Measures AA1 and AA2, providers who operate a single service may be particularly disadvantaged relative to providers who can offer access to a range of services with a single age check. However, we consider that this effect is likely to be more limited with Measures AA3 and AA4 than Measures AA1 and AA2, given that services can allow adults to continue to use the service without access to identified PPC and PC content.

Which providers we propose should implement these measures

15.204 As noted above, Measure AA3 applies to all services that do not prohibit all kinds of PPC (and are not already in scope of Measure AA1), which closely reflects the specific requirement under the Act for these kinds of services over which we have no discretion. As such, we consider that Measure AA3 is proportionate as applied to those services.

15.205 We exercise discretion in proposing Measure AA4, which applies where the service does not prohibit all kinds of PC (and is not already in scope of Measure AA2), but with the added

condition that the service is medium or high risk for at least one kind of PC that it allows. In other words, if a service has low risk for all kinds of PC – despite not prohibiting all kinds of PC – then it would not be in scope of Measure AA4. This reflects that the benefits of the measure, from increased protection of children from PC, would be limited for such a service, whereas the costs are still likely to be material as described above.

15.206 We have also considered whether the proposed measure is proportionate for smaller services that have only medium risk for one kind of PC. As set out in the effectiveness sub-section, we consider that exposure to only one kind of PC can still cause severe harm to children, including in cases of repeated exposure, and we consider this outcome inherently more likely on services that do not prohibit such content. Such harm can occur on smaller services as well as large ones. Therefore, we believe that the proposed measure targeted at services who pose such risks can support significant incremental benefits to children, even on smaller services with a single risk of PC.

15.207 We are not currently proposing to recommend this measure for services based on whether or not they prohibit NDC. This is because we have less evidence at this stage about services and harm associated with NDC.

15.208 Unlike Measures AA1 and AA2, services are free to apply highly effective age assurance only to those users who seek to access identified PC, which can limit costs. Nonetheless, the costs associated with our measures can be material and could mean some services choose to stop serving PC to UK users, which would affect the ability of adult users to access this type of content. Some services may decide to stop serving the UK altogether, which would limit the ability for adult users to access PC on these services, and for child users to benefit from the non-harmful content on these services. Other services may choose to use highly effective age assurance to prevent children's access to the whole service, which would prevent them accessing non-harmful and beneficial content on such services. These examples are not the intended effects of the proposed measure. We think services should instead find ways to create age-appropriate inclusive environments that allow children to enjoy the benefits of this technology while protecting them from harm. While some smaller services may not be able to achieve this, we believe that the flexibility we allow for services in terms of how to implement highly effective age assurance should ensure many will, meaning that PC will remain accessible to adults on a wide range of services.

15.209 As is the case for Measures AA1 and AA2, our measures will make it more cumbersome for adults to access PC on services that allow it and pose a risk of it to children. While services may reduce this "hassle factor" by requiring a one-off age check associated with an account, this can reduce the ability for users to access this content without being logged in. This could have privacy impacts or limit adults' freedom of expression by dissuading them from accessing PC altogether. We also acknowledge that children who are not necessarily severely harmed by PC are nonetheless prevented from accessing this content, impacting their freedom of expression. However, as explained in previous sub-sections, we consider these measures to be the only feasible way to secure providers of these kinds of services comply with the duties set out in the Act, with clear potential to substantially improve children's safety online even in respect of smaller services.

15.210 We have also considered whether the measure should apply for all types of U2U services, including messaging services (as discussed also in the 'Privacy' sub-section above). We provisionally conclude that this is proportionate, as we believe material harm from encountering PC can occur across a broad range of services, including services with one-to-

one or group messaging functionalities. However, we recognise there may be certain challenges in taking effective actions to protect children from encountering harmful content via private communication channels and we welcome input on this issue in consultation responses.

15.211 In summary, we propose that Measure AA4 applies to services that:

- Are not in scope of Measure AA2 – that is, their principal purpose is not the hosting or dissemination of one of more kinds of PC;
- Do not prohibit one or more kinds of PC; and
- Have medium or high risk for one or more kinds of PC that they do not prohibit.

Provisional conclusion

15.212 Given the harms these measures seek to mitigate in respect of PPC and PC, we consider these measures appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for these measures, please see PCU H4 and H5 in Annex A7. Please also see Annex 10 (draft HEAA guidance).

Targeting recommender system measures

15.213 Age assurance can enable the targeting of safety measures to children. Our proposed recommender systems measures rely on services determining which users are children to tailor their recommender feeds appropriately.

15.214 In summary, the recommender system measures we propose are linked to our proposed age assurance measures as follows (see Section 20 for more detail):

a) Recommender Systems Measure RS1 involves filtering out content likely to be PPC from recommender feeds of children. Our proposed Measure AA5 recommends the use of highly effective age assurance for the purpose of targeting Recommender Systems Measure RS1.
b) Recommender Systems Measure RS2 involves reducing the prominence of content likely to be PC from recommender feeds of children. Our proposed Measure AA6 recommends the use of highly effective age assurance for the purpose of targeting Recommender Systems Measure RS2.
c) Recommender Systems Measure RS3 involves providing children with a means of expressing negative sentiment to provide negative feedback directly to their recommender feed. If a service is in scope of this measure, it will also be in scope of our proposed Measure AA5 and/or Measure AA6, and it may use highly effective age assurance also for the purpose of targeting Recommender Systems Measure RS3.

Measure AA5: Use HEAA to apply relevant recommender system measures to protect children from PPC

Services that are high or medium risk for one or more kinds of PPC and operate a recommender system, should use highly effective age assurance to apply the relevant recommender system measures in the Code to children.

Explanation of the measure

15.215 Age assurance can enable further safety measures to work in a more targeted and effective way. Once services establish that a user is a child, they can direct safety measures to them which offer an additional layer of protection. Measure AA5 is designed to introduce highly effective age assurance for the purposes of targeting recommender system safety measures to children. We have set out above at sub-section 'Age assurance for other protection of children measures' how age assurance can similarly be used for the targeting of user support measures under the children's safety duties, and measures to prevent children from encountering illegal harms such as grooming. We expect age assurance will be important in targeting additional protections in future.

15.216 Proposed Measure AA5 will apply to services who are high or medium risk for one or more kinds of PPC (whose principal purpose is not to host or disseminate PPC) and have a recommender system.[134] Services in scope of Measure AA3 who have a recommender system are also in scope of Measure AA5, and would be recommended to implement highly effective age assurance for the purposes of both measures.

15.217 Services would be in scope of Measure AA5 if they are also in scope of the proposed Recommender Systems Measure RS1 set out in Section 20 (Recommender systems on U2U services), which provides that services should design their recommender systems to filter out content that is likely to be PPC from children's recommender feeds. Recommender system Measure RS1 relies on services identifying children accurately. We discuss the workings of these measures in more detail in Section 20 (Recommender systems on U2U services).

15.218 In developing our proposals for age assurance we considered the mechanisms by which children may encounter harmful content. Given the nature of recommender systems, in that they may increase the risk of children encountering harmful content even without their actively seeking or engaging with it, and risk cumulative exposure leading to material harm, we believe these systems materially influence what users, and therefore children, see. Furthermore, services that generate revenue in proportion to user engagement (for instance, advertising revenue models and subscription revenue models) can have incentives to develop service designs and features that maximise engagement and drive revenue at the expense of exposing child users to harmful content. This measure provides a route to reduce children's exposure to such content. Without understanding whether a user is a child, it is

[134] The measures proposed in the recommender systems section would only apply to content recommender systems, and not to recommender systems that underpin search functionalities on a U2U service, or network recommender systems that suggest other users to follow or groups to join. For definitions, please refer to Section 20.

not possible for services to implement our proposed measures for recommender systems effectively. For this reason, we propose to recommend the use of highly effective age assurance combined with the proposed recommender system safety measures to ensure they are applied to users correctly.

15.219 To implement Measure AA5 and ensure that the Recommender Systems Measure RS1 in Section 20 applies correctly to children, providers should use age assurance of such a kind, and in such a way, that is highly effective at correctly determining whether or not a particular user is a child. Our draft guidance on how services should implement highly effective age assurance is set out in Annex 10.

15.220 It is for the service provider to determine how to implement a highly effective age assurance process on its service to secure these outcomes. To do so, the service provider must ensure that at whatever point users are required to undergo the age check, children have the relevant restrictions for content that is likely to be PPC on their recommender feeds. One way of doing this would be to implement age assurance at the point where users first access the service, so that content that is likely to be PPC can be filtered out of all recommender feeds for those not determined to be adults.

15.221 However, there may be alternative ways of achieving this outcome. For example, services might offer users the option to unlock an unfiltered recommender feed by conducting an age check, without necessarily implementing age assurance for *all* users accessing the service. The key point is that services must secure that all users who may be children (i.e., are not determined to be adults, which would include logged-out users who have not undergone any form of age assurance) have content likely to be PPC filtered out of their recommender feeds (see Section 20 recommender systems on U2U services).

15.222 In deciding when and how to implement highly effective age assurance, we encourage service providers to consider the potential impacts on users, for instance how their proposed approach might affect the level of friction experienced by users. Regardless of how the measure is implemented, it should secure the outcome that content likely to be PPC is consistently filtered out of the recommender systems for children.

Effectiveness at addressing risks to children

15.223 The Act deems PPC to be harmful for children and sets out that children must be prevented from encountering it.

15.224 In Section 20 (recommender systems on U2U services) we recommend measures to prevent children's exposure to PPC content via recommender systems. We discuss the risks to children presented by recommender systems, and how our proposed measures address these risks in detail at Volume 3, Section 7.11 Governance, systems and processes.

15.225 We consider that Measure AA5 will enable the recommender systems measures to effectively address the risk of children encountering content likely to be PPC on recommender feeds. It will do so by enabling services to distinguish between adults and children using highly effective age assurance to accurately target those recommender system safety measures towards children while allowing adults to view unfiltered recommendations.

15.226 This includes protecting potential child users in logged out environments. The approach of seeking to create a safe environment for all logged out users, who would be non-age assured users, is in line with industry practice where services claim to offer logged out users an

environment that is suitable for all users. For example, our 2022 VSP report found that Vimeo does not allow users without an account to view videos that are rated mature or left unrated to reduce the likelihood of children encountering harmful content.[135] Similarly, content that is available to logged out users on TikTok undergoes several rounds of human moderation, and any content that has a warning notice or videos with captions or hashtags that hit 'sensitive word lists' will not be eligible to appear to users not logged in.[136] Snap also uses this approach on the Discover and Spotlight feed. Content such as sexually explicit content, violence, or dangerous behaviour, is prohibited on Discover and Spotlight by Snap's Community Guidelines and is reviewed by human moderators and automatic content moderation tools.[137] Whilst these current practices may not sufficiently create safe environments for users, they demonstrate that services already have some measures in place that aim to achieve an experience for logged out users which differs from the logged in experience of users.

15.227 We considered whether a form of age assurance offering a lower level of assurance could be sufficient to secure the outcomes aimed at by this measure, namely, to ensure that content that is likely to be PPC is filtered out of children's recommender feeds. We do not consider this to be the case as a higher proportion of children accessing the service might not end up benefiting from these protections. We discuss our considerations of alternatives to highly effective age assurance, and our rationale for not recommending these, further at sub-sections 'Other options considered.'

Rights assessment

15.228 This proposed measure will apply to those services at high or medium risk of one or more kinds of PPC, whose principal purpose is not the hosting or dissemination of PPC and that use a recommender system(s). It is intended to support the proposed measures in Section 20 (recommender systems for U2U services) so that services can apply those safety recommendations appropriately. By preventing children's access to PPC, the proposed measure will seek to secure adequate protections for children from harm, in line with the legitimate aims of the Act. Preventing children from encountering PPC acts to prevent the harmful consequences of such content that can be inflicted on them. These consequences can include harm to children's physical, mental or emotional wellbeing. The proposal does not mandate a specific method of age assurance and is designed to follow our criteria-based approach to implementing highly effective age assurance.

Freedom of expression and association

15.229 We consider that this proposed measure has the potential to impact users' (both adults' and children's) rights to freedom of expression and of association as also set out in relation to the measures detailed above.

15.230 As with Measures AA1 and AA3, the duty on services that do not prohibit PPC to use highly effective age assurance as part of their systems and processes to prevent children from encountering PPC identified by the service is a requirement of the Act. To the extent that the application of highly effective age assurance effectively prevents children from encountering PPC and adults' and other users' ability to share such content with children in their

[135] Ofcom, 2022. Ofcom's first year of video-sharing platform regulation.
[136] Ofcom, 2023. How video-sharing platforms (VSPs) protect children from encountering harmful videos.
[137] Ofcom, 2023. How video-sharing platforms (VSPs) protect children from encountering harmful videos.

recommender feeds, we consider that this is justified and proportionate in line with the duties of the Act.

15.231 Many of the potential impacts of this measure on children's and adult users' rights to freedom of expression and association arise as a result of the way that Recommender Systems Measure RS1 in Section 20 has been designed. For example, the fact that it applies to content likely to be PPC and not just content identified as PPC, which means there is a potential risk that there may be cases where content that is not PPC, including content that may not be harmful to children, is flagged as likely to be PPC and removed from children's recommender feeds as a result of this measure (for example, due to inaccurate labelling). We have explained why we consider the impact of filtering out content likely to be PPC from children's recommender feeds to be proportionate to the potential limited impacts this has on users' and service providers' rights to freedom of expression, given the way we have designed this proposed measure, in Section 20 and we do not repeat these here.

15.232 We recognise that there might be additional impacts on adult users' rights to freedom of expression if adult users are wrongly determined not to be adults and this means that they are unable to access content likely to be PPC in their recommender feeds as a result, particularly where such content in fact is not PPC. However, as noted in connection with the above measures, we consider this risk to be limited provided services implement our recommended principles for highly effective age assurance, and also the complaints process that services will be required to make available as per Section 18 (user reporting and complaints).

15.233 Other than what we have outlined above, we consider that the potential impacts on adult users' rights to freedom of expression as a result of this measure are similar to those outlined in relation to Measure AA3 above for those services that are also in scope of Measure AA3 because they do not prohibit one or more forms of PPC. This includes the potential impacts which could arise if services choose to withdraw their recommender system, or to withdraw the service from the UK market entirely (for instance, if the recommender system is integral to the service's business model) due to the costs of implementing highly effective age assurance, together with the cost of implementing recommender system changes under the related Recommender Systems Measure RS1 in Section 20.

15.234 For those services that are in scope of this measure, but are not in scope of Measure AA3 above because they prohibit all forms of PPC for all users, we do not consider there are any additional relevant freedom of expression impacts other than those already discussed above. This is because if content likely to be PPC identified via this measure is confirmed as PPC, it would need to be removed from the service in any case as it would violate their terms of service, and if it was ultimately not found to be violative, it could be reinstated so that all users, including children, could then access it in their recommender feeds (see Section 20).

15.235 For the reasons set out above and in Section 20, we consider that the impact of the proposed measure on users' and service providers' rights to freedom of expression to be limited, and no further than needed to secure the positive benefits to children in preventing their exposure to PPC, in line with the requirements of the Act. We consider that is therefore proportionate.

Privacy

15.236 We consider that this proposed measure has the potential to impact on users' (both adults' and children's rights) to privacy for the reasons set out in relation to Measures AA1 to AA4 above.

15.237 As set out in Measures AA1 to AA4 above, all age assurance processes will inevitably involve the processing of personal data of individuals, including children. There are particular risks in relation to privacy and personal data if more personal data than needed is processed as a result of the age assurance process, or if users' ages are incorrectly assessed, for example adult users prevented from being recommended this content, or children encountering this content if they are incorrectly assessed as adult users. This could result in services (and third-party age assurance providers) having more personal data than needed or inaccurate personal data of users. We consider this risk can be mitigated by services having place appropriate complaints policies and processes as set out in our Section 18 (user reporting and complaints). Services will also need to comply with data protection laws and ICO guidance, as set out in Measures AA1 to AA4 above to ensure that users are able to fully exercise their rights in respect of their personal data.

15.238 We therefore consider that the impact of the proposed measure because of services' implementation of highly effective age assurance on child and adult users' rights to privacy, to be potentially significant. However, we have not identified any specific potential impacts connected with restrictions on children's or adults' private communications, unlike in respect of Measure AA3 above, as by their nature, recommender systems would generally only promote content that is widely publicly available, rather than private communications. In this respect, we consider the impact of this proposed measure to be more limited.

15.239 Assuming service providers also comply with data protection legislation requirements, our provisional view is that the degree of interference with users' rights to privacy as a result of this measure is likely to go no further than needed to secure the positive benefits to children in preventing their exposure to PPC, in line with the requirements of the Act. Taking this, and the significant benefits to children into consideration, we consider that the interference with users' rights to privacy is therefore proportionate.

Measure AA6: Use HEAA to apply relevant recommender system measures to protect children from PC

Services that are high or medium risk for one or more kinds of relevant PC and operate a recommender system, should use highly effective age assurance to apply the relevant recommender system measures in the Code to children.

Explanation of the measure

15.240 Measure AA6 is also designed to introduce highly effective age assurance for the purposes of targeting recommender system safety measures to children. Once services can accurately determine which users are children using age assurance, they can direct safety measures towards them to keep them safe online.

15.241 Measure AA6 will apply to those services in scope of the proposed Recommender Systems Measure RS2 set out in Section 20. This recommends that services with recommender systems that are high or medium risk for any relevant kinds of PC significantly limit the prominence and visibility of content that is likely to be PC in children's recommender feeds. Services in scope of Measure AA4 who have a recommender system are also in scope of Measure AA6, although for such services, we consider that Measures AA4 and AA6 would both require the same outcomes in respect of protecting children from encountering content likely to be PC via their recommender feeds.

15.242 We are proposing not to include bullying content under "relevant kinds of PC" for this measure. This is because there is insufficient evidence that recommender systems are a contributing factor in the exposure of children to bullying content. This is explained further in Section 20 on Recommender Systems.

15.243 Subject to the outcome of the consultation on NDC, we are also minded to recommend that body image and depressive content (NDC) is included within measure RS2, which recommends that services limit the prominence of this content. Under such circumstances, we would also be minded to propose that services with body image and depressive content also fall within scope of measure AA6. This is explained further in Section 20. We discuss the rationale for proposing to recommend the use of highly effective age assurance in parallel with the proposed recommender system safety measures specifically in Measure AA5 above.

15.244 To comply with Measure AA6 and secure the outcome that Recommender Systems Measure RS2 applies correctly to children on their service, providers should use age assurance of such a kind, and in such a way, that is highly effective at correctly determining whether or not a particular user is a child. Our draft guidance on how services should implement highly effective age assurance is set out in Annex 10.

15.245 It is for providers to determine how to implement highly effective age assurance on the service to secure these outcomes. To do so, the service provider must ensure that at whatever point users are required to undergo the age check, children have the relevant restrictions for content that is likely to be PC applied to their recommender feeds.

15.246 One way of doing this would be to implement age assurance for users at the point where users first access the service so that the prominence of content likely to be PC can then be significantly limited in all recommender feeds for those not been determined to be adults.

15.247 However, there may be alternative ways of achieving this outcome. For example, services might offer users the option to unlock a recommender feed without the safety measure applied, by conducting an age check, without necessarily implementing age assurance for *all* users accessing the service. The important point is that services must secure that all users who may be children (i.e., are not determined to be adults, which would include logged-out users who have not undergone any form of age assurance) have content likely to be PC significantly limited in their recommender feeds. As set out at in measure AA5, the approach of seeking to create a safe environment for all logged-out users who have not been age assured mirrors current industry practice.

15.248 In deciding how to implement the measure, service providers may want to consider the potential impacts on users, for instance, how its proposed approach to implementing highly effective age assurance might affect the user experience of the service. Regardless of how the measure is implemented, it should secure the outcome that the prominence of content that is likely to be PC is significantly limited and therefore is less visible to children in their recommender feeds.

Effectiveness at addressing risks to children

15.249 The Act requires that children in age groups judged to be at risk of harm from PC are protected from encountering PC. We discuss the wide-ranging negative impacts on children of encountering PC in detail in Volume 3 of this consultation.

15.250 We recommend specific measures to protect children from exposure to PC content in their recommender feeds in Section 20. The recommender system measures will address the risk of those systems perpetuating PC harm by applying filters to children's recommendations to significantly limit the prominence of content that is likely to be PC. We explain how those measures will effectively address this risk in more detail in Section 20.

15.251 Implementing highly effective age assurance under Measure AA6 will enable services to distinguish between adults and children, to accurately target those recommender system safety measures towards children, while allowing adults to view unaltered recommendations.

15.252 We considered whether a lower level of assurance would be more proportionate for this measure to reflect that the relevant duty on services is to "protect" children from encountering PC, rather than to "prevent" children from encountering PC as under Measure AA3. We have discussed the problems we encountered in defining a lower level of age assurance in more detail at sub-section 'Other options considered.' Ultimately, we concluded that the recommender systems measures have already been designed in such a way to reflect this difference, in that services should significantly limit the prominence of content that is likely to be PC in recommender feeds but filter out content that is likely to PPC entirely in the recommender feeds of children.

15.253 We also considered whether it would be possible to recommend that services tailor this measure so that access to the service would only be prevented by age groups judged to be at risk of harm, as identified in the service's children's risk assessment. However, we currently have limited evidence linking specific PC harms to different age groups. We will continue to review this and discuss this more in our 'Children in different age groups' sub-section of this section and welcome answers to question 4 in our 'Consultation questions' in this section.

Rights assessment

15.254 This proposed measure recommends that all services in scope of proposed Recommender Systems Measure RS2 in Section 20 use highly effective age assurance to determine which users are children. No specific method of age assurance is mandated within this measure in line with our criteria-based approach.

15.255 The proposal is intended to support that proposed measure so that services can apply those safety recommendations appropriately. By protecting children from encountering PC, the proposed measure will seek to secure adequate protections for children from harm, in line with the legitimate aims of the Act. Preventing children from encountering PC acts to prevent the harmful consequences of such content that can be inflicted on them. These consequences can include harm to children's physical, mental or emotional wellbeing.

Freedom of expression and association

15.256 The proposed measure does not recommend services restrict access to adult users but instead it seeks to secure that the provider's systems or processes are designed so that they take steps to protect children from encountering PC. However, we consider that this

proposed measure has the potential to significantly impact on users' (both adults' and children's rights) to freedom of expression and of association for the reasons set out in relation to Measures AA2 and AA4 above. To the extent that the application of highly effective age assurance effectively protects children from encountering PC and adults' and other users' ability to share such content with children in their recommender feeds, we consider that this is justified and proportionate in line with the duties of the Act.

15.257 Many of the potential impacts of this measure on children's and adult users' rights to freedom of expression and association arise as a result of the way that Recommender Systems Measure RS2 in Section 20 has been designed. For example, the fact that it applies to content likely to be PC and not just content identified as PC, which means there is a potential risk that there may be cases where content that is not PC, including content that may not be harmful to children, is flagged as likely to be PC and removed from children's recommender feeds as a result of this measure (for example, due to inaccurate labelling). We have explained why we consider that the impact of limiting visibility of content likely to be PC from children's recommender feeds to be proportionate to the potential limited impacts this has on users' and service providers' rights to freedom of expression, given the way we have designed this proposed measure, in the Recommender Systems Section 20 and we do not repeat this discussion here.

15.258 We recognise that there might be additional impacts on adult users' rights to freedom of expression if adult users are wrongly identified as child users and this means that they are unable to access some content likely to be PC in their recommender feeds as a result, particularly where such content in fact is not PC. However, as noted in connection with the above measures, we consider this risk to be limited provided services implement our recommended principles for highly effective age assurance, and also the complaints process that services will be required to make available as per Section 18 (user reporting and complaints).

15.259 We consider that there may be additional impacts on adult users' rights if services choose to remove all content likely to be PC from all users' recommender feeds but not other parts of the service, so as to potentially avoid the costs of applying highly effective age assurance – but this is not something we are expressly recommending and it remains services' commercial choice as to what forms of content to allow to be made available on what parts of the service.

15.260 Otherwise, we consider that the potential impacts on adult users' rights to freedom of expression as a result of this measure are the same as those outlined in relation to Measure AA4 above for those services that are also in scope of Measure AA4 because they do not prohibit one or more forms of PC and are medium or high risk of those forms of PC appearing. This includes the potential impacts which could arise if services choose to withdraw from the UK market due to the costs of implementing highly effective age assurance (for instance, if the recommender system is integral to the services' business model), and the negative impacts on the experience of older children in age groups not judged at risk of harm from the relevant forms of PC who would be unable to access such PC as freely via their recommender feeds, though we consider the latter impact to be more limited to the extent that it only involves less visibility of PC, rather than stricter access or content controls as may result from some implementations of Measure AA4.

15.261 For those services that are in scope of this measure, but are not in scope of Measure AA4 above because they prohibit all forms of PC for all users, we do not consider there are any

additional relevant freedom of expression impacts other than those already discussed above. This is because if content likely to be PC identified via this measure is confirmed as PC, it would need to be removed from the service in any case as it would violate their terms of service, and if it was ultimately not found to be violative, it could be reinstated so that all users, including children, could then access it in their recommender feeds.

15.262 For the reasons set out above, and in Section 20 on Recommender Systems, we consider that the impact of the proposed measure on users' and service providers' rights to freedom of expression to be relatively limited, and taking into account the significant benefits of protecting children from harm arising from PC which may otherwise occur, our provisional view is that the interference with users' and service providers' rights to freedom of expression and association is therefore proportionate.

Privacy

15.263 We consider that this proposed measure has the potential to impact on users' (both adults' and children's) rights to privacy for the reasons set out in relation to Measures AA1 to AA5 above, as all methods of age assurance will involve the processing of personal data of individuals, including children, whose personal data requires special consideration.

15.264 However, we have not identified any specific potential impacts connected with restrictions on children's or adults' private communications, unlike in respect of Measure AA4 above, as by their nature, recommender systems would generally only promote content that is widely publicly available, rather than private communications. In this respect, we consider the impact of this proposed measure to be more limited.

15.265 Assuming service providers also comply with data protection legislation requirements, our provisional view is that the degree of interference with users' rights to privacy as a result of this measure is likely to go no further than needed to secure the positive benefits to children in protecting them from PC, in line with the requirements of the Act. Taking this, and the significant benefits to children into consideration, we consider that the interference with users' rights to privacy is therefore proportionate.

Impacts on services – Measures AA5 and AA6

15.266 We expect the cost of implementing highly effective age assurance under Measures AA5 and AA6 to be similar to those discussed in more detail under Measures AA1 and AA2 above. Costs of age assurance can be substantial and are likely to depend on the approach a service takes to implementation (e.g., the method(s) used), and the number of users service has.

15.267 However, whereas Measures AA1 and AA2 would mean implementing highly effective age assurance for every user accessing the service, Measures AA5 and AA6 may only require age checks for a subset of users. Similarly to our approach in Measures AA3 and AA4, with Measures AA5 and AA6 we provide discretion for services to determine at which point of the user journey they implement age assurance. For example, a service might decide to implement by offering the option to conduct an age check for adult users who specifically want to see PPC and PC in their recommender feeds, and applying the recommender system safety measures to all other users. This could limit the costs, depending on the number of users who choose to undertake an age check, which will vary depending on the nature of the service and its user base.

15.268 As well as the costs of implementing age assurance, providers may incur substantial costs to implement the recommender system measures themselves and ensure that the right

protections are applied to users based on whether they have conducted an age check as well as the outcome of any age check. The costs of making the relevant recommender system changes are discussed separately in Section 20.

15.269 Services may also experience indirect costs through reduced user engagement, user numbers and revenue if there are adults who are discouraged from using the service because they want to access PPC and PC in their recommender feeds, but they are unwilling to complete the age assurance process. As discussed for Measures AA1 and AA2, providers who operate a single service may be particularly disadvantaged relative to providers who can offer access to a range of services with a single age check. However, we consider that this effect is likely to be more limited with Measures AA5 and AA6 than, for example, Measures AA1 and AA2, given that services have the option to allow adults to continue accessing recommender feeds with safety measures applied but still including a wide variety of content that can be appealing to adults that is not PPC or PC.

Which providers we propose should implement these measures

15.270 We propose that services that operate a recommender system and are high or medium risk of any kind of PPC or any relevant kind of PC, should introduce highly effective age assurance to align the scope of our access and content control measures with those of the Recommender Systems Measures RS1 and RS2 to achieve the intended effect. In Section 20 we have considered the impacts and proportionality of these measures, including the kinds of services that the measures should apply to, taking also into account the implications of implementing Measures AA5 and AA6 as described here.

15.271 We think that the use of highly effective age assurance is necessary to ensure that children receive the intended recommender system measure protections. Where a service has a recommender system in place, there is an increased risk of children encountering PPC and/or PC even without them actively seeking or engaging with it, and there are significant limitations of effectiveness if these measures are applied only based on self-declaration. Therefore, we propose that all user-to-user services that have a recommender system and are high or medium risk for any kinds of PPC and/or any relevant kinds of PC should use highly effective age assurance to correctly apply the relevant recommender system measures as discussed in Section 20 to children.

15.272 Unlike Measures AA1 and AA2, services are free to apply highly effective age assurance only to those users who seek to access likely PPC or PC by means of a recommender system, which can limit costs. Nonetheless, the costs associated with our Measures AA5 and AA6 can be material and could lead some services to withdraw their recommender system functionality or consider it too expensive to serve UK online users because of our proposed measures. This would adversely impact adult and child users. These examples are not the intended effects of the proposed measure. We think services should instead find ways to create age-appropriate inclusive environments that allow children to enjoy the benefits of this technology while protecting them from harm. While some smaller services may not be able to achieve this, we believe that the flexibility we allow for services in terms of how to implement highly effective age assurance should ensure many will, meaning that recommender systems will remain available to UK users.

15.273 Our measures will make it more cumbersome for adults to access PPC and PC via recommender systems. While services may reduce this hassle factor by requiring a one-off age check associated with an account, this can reduce the ability for users to access this

content without being logged in. This could adversely affect the user experience of adults and impact their privacy. However, as explained in previous sub-sections, we consider these measures to be the only feasible way to secure providers of these kinds of services comply with the duties set out in the Act, with clear potential to substantially improve children's safety online even in respect of smaller services.

15.274 In summary, given the risks of harm to children posed by recommender systems, we believe these measures to be proportionate when applied to all U2U services that have a recommender system and are high or medium risk for any kinds of PPC or relevant kinds of PC.

Provisional conclusion

15.275 Given the harms measures AA6 and AA7 seek to mitigate in respect of PPC and relevant kinds of PC, we consider these proposed measures appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for these measures, please see PCU H6 and H7 in Annex 7. Please also see Annex 10 (draft HEAA guidance).

Our approach to highly effective age assurance

How we developed our approach in the Codes

Consistency with Part 5

15.276 In December 2023, we published a consultation on our draft guidance for service providers publishing pornographic content under Part 5 of the Act ('Part 5 Consultation'). The Part 5 duties require Part 5 service providers to implement age assurance to ensure that children are not normally able to encounter regulated provider pornographic content on the service.[138] The age assurance used must be of such a kind, and used in such a way, that is highly effective at correctly determining whether or not a particular user is a child.[139] Ofcom is required to publish guidance to assist services in complying with the age assurance duties.[140]

15.277 In developing our guidance, we considered whether to specify a numerical threshold for accuracy that the age assurance method(s) should achieve to be considered highly effective. Given the evidence available to us at that time, and the developing nature of the age assurance industry, we decided instead to propose a set of criteria that service providers should ensure their age assurance method(s) or processes fulfils to be highly effective. We sought views and evidence from stakeholders on this approach through consultation.

15.278 To maintain consistency with the Part 5 Consultation, as well as the approach to age assurance that we set out as part of our Section 4 children's access assessments, we are proposing a criteria-based approach to highly effective age assurance under the age assurance measures. These criteria are technical accuracy, robustness, reliability and

[138] Section 81(2) of the Act. 'Regulated provider pornographic content' refers to pornographic content that is published or displayed on the service by the provider of the service or by a person acting on behalf of the provider (section 79(2) of the Act).
[139] Section 81(3) of the Act.
[140] Section 82 of the Act.

fairness and are discussed further in Annex 10 (draft HEAA guidance) and also in sub-section 'Proposed draft Code Measures' below.

15.279 We are carefully considering our position in light of the views and evidence provided by stakeholders as part of the Part 5 consultation process. We will seek to maintain consistency in our approach to highly effective age assurance in developing our final guidance and Codes across both Part 3 and Part 5 of the Act.[141]

Schedule 4 principles

15.280 We have also considered the principles set out in Schedule 4 of the Act in developing both our measures, and our recommendations around the types of age assurance that our measures rely on.

15.281 As explained in sub-section 'Our proposals to protect children', we have taken into account the principle that "more effective kinds of age assurance should be used to deal with higher levels of risk of harm to children," taking into account the nature and severity of potential harm to children[142], in developing Measures AA1 to AA6 which explain when and how highly effective age assurance should be used to prevent children from encountering PPC or protect them from encountering PC.

15.282 We have also taken into account relevant Schedule 4 principles in developing our approach to highly effective age assurance to ensure accessibility, for adults and children, and effectiveness for all users regardless of their characteristics. In addition, we have considered the principle of interoperability between different kind of age assurance (where possible).[143]

15.283 Figure 15.3 below sets out where we have considered each of the principles set out in Schedule 4 of the Act in developing our recommendations and guidance on highly effective age assurance.

[141] Dedicated pornography services may fall under Part 3 and/or Part 5 of the Act. Where the majority of the content hosted by a dedicated pornography service is user-generated pornographic content, the service should fall in scope of age assurance Measure 1. Where a service hosts provider pornographic content it should fall under Part 5 of the Act.
[142] Schedule 4, Paragraph 12 of the Act.
[143] Schedule 4, Paragraph 12(2) of the Act.

Figure 15.3: Considerations in relation to Schedule 4 principles.

Schedule 4 principle	Consideration
The principle that age assurance should be effective at correctly identifying the age or age-range of users.	We recommend criteria that service providers should ensure the age assurance process fulfils to be highly effective at correctly identifying the age or age-range of users.
Relevant standards set out in the latest version of the code of practice under Section 123 of the Data Protection Act 2018 (age-appropriate design code).	We recommend that when implementing age assurance, services familiarise themselves with the standards in the ICO Children's code, and the Commissioner's Opinion on Age Assurance for the Children's code (the Opinion). See sub-section 'Privacy and data protection' in the draft HEAA guidance on our approach to privacy and data protection. We have provided examples of how services can demonstrate consideration of data protection laws, drawing in particular on Standards 2 and 4, and the Governance and Accountability guidance provided in the Children's code. See sub-section 'Privacy and data protection' of the draft HEAA guidance for additional discussion of Standard 4 under the Children's code and how it relates to highly effective age assurance.
The need to strike the right balance between – The levels of risk and the nature, and severity, of potential harm to children which the age assurance is designed to guard against, and, Protecting the right of users and interested persons to freedom of expression	We discuss the rights impacts of each of our proposed measures in the 'rights assessment' sub-sections of AA1-AA6.

Schedule 4 principle	Consideration
The principle that more effective kinds of age assurance should be used to deal with higher levels of risk of harm to children	See 'other options considered' where we considered alternative forms of age assurance, and have proposed that highly effective age assurance should be used in cases where: • A service does not prohibit one or more kinds of PPC (whether or not the hosting or dissemination of PPC is its principal purposes). • A service does not prohibit one or more kinds of PC and is medium or high risk for that kind of PC (whether or not the hosting or dissemination of PC is its principal purpose) • A service is medium or high risk of one or more kinds of PPC or PC and has a recommender system (whether or not it prohibits that kind of PPC or PC).
The principle that age assurance should be easy to use, including by children of different ages and with different needs	We have included this principle in our recommendations on highly effective age assurance. We also provide guidance on this in the 'Transparency' section of the draft HEAA guidance.
The principle that age assurance should work effectively for all users regardless of their characteristics or whether they are members of a certain group	We have included this principle in our recommendations on highly effective age assurance. We also provide guidance on this in sub-section 'Accessibility' in our draft HEAA guidance.
The principle of interoperability between different kinds of age assurance.	We have included this principle in our recommendations on highly effective age assurance. We provide guidance on interoperability in our draft HEAA guidance.

Proposed draft Code measure

15.284 We are proposing to set out in the Children's Safety Codes of Practice the following measure for U2U services in scope of measures AA1 to AA6 above to describe the meaning of highly effective age assurance for those measures.

15.285 For the use of age assurance to be highly effective at correctly determining the age of users, service providers should choose an appropriate method (or methods) of age assurance that is of such a kind that could be highly effective at correctly determining whether a user is a child.

15.286 Service providers should ensure that their chosen age assurance process as a whole fulfils each of the criteria of technical accuracy, robustness, reliability and fairness, to ensure it is highly effective in practice.

15.287 The technical accuracy criterion is fulfilled if:

a) the provider has ensured that the measures[144] forming part of the age assurance process for the service have been evaluated against appropriate metrics to assess the extent to which they can correctly determine the age or age range of a person under test lab conditions;
b) where the age assurance process used on the service involves the use of age estimation, the provider uses a challenge age approach; and
c) the provider periodically reviews whether the technical accuracy of the age assurance process for the service could be improved by making use of new technology and, where appropriate, makes changes to the age assurance process.

15.288 The robustness criterion is fulfilled if:

a) The provider has:

i) taken steps to identify methods children use to circumvent the age assurance process used on the service to determine that the relevant individual is not a child; and
ii) taken feasible and proportionate steps to prevent children using those methods; and

b) the provider has ensured that the age assurance measures forming part of the age assurance process for the service have been tested in multiple different environments during the development of the age assurance process.

15.289 The reliability criterion is fulfilled if:

a) where age assurance measures forming part of the age assurance process rely on artificial intelligence or machine learning, the provider has taken steps to ensure that:

i) the artificial intelligence or machine learning has been suitably tested during the development of the age assurance process to ensure it produces reproducible results;
ii) the artificial intelligence or machine learning is regularly tested to ensure it produces reproducible results;
iii) the outputs of the artificial intelligence or machine learning used are monitored and assessed against key performance indicators designed to identify whether the artificial intelligence or machine learning produces reproducible results;
iv) in circumstances where the artificial intelligence or machine learning used are observed to be producing unreliable or unexpected results, the root cause of the issue is identified and rectified.

b) The provider has taken steps to ensure that any data relied upon as part of the age assurance process comes from a reliable source.

---

[144] We acknowledge the Draft Code Children Safety Codes refers to 'age assurance methods' as 'age assurance measures.' This is to reflect the statutory language of the Act as per Section 41(3), Schedule 4, in particular Sch 4 para 12. In sub-section 'Highly Effective Age Assurance' of the Draft Children Safety Codes, 'age assurance measures' has the same meaning as ‘age assurance methods’ in the Age Assurance Section.

15.290 The fairness criterion is fulfilled if:

a) The provider has ensured that any elements of the age assurance process for a service, which rely on artificial intelligence or machine learning have been tested and trained on data sets which reflect the diversity in the target population.

15.291 Service providers should not publish content that directs or encourages United Kingdom users to circumvent the age assurance process or access controls used on the service.

15.292 When implementing the age assurance process, service providers should have regard to the following principles:

- the principle that age assurance should be easy to use, including by children of different ages and with different needs;
- the principle that age assurance should work effectively for all users regardless of their characteristics or whether they are members of a certain group;
- the desirability of ensuring interoperability between different kinds of age assurance;
- the latest version of the age appropriate design code and the Information Commissioner's opinion entitled "Age Assurance for the Children's code" published on 18 January 2024.

15.293 The provider should ensure that users are able to easily access information about what a provider's age assurance process is intended to do and how the provider's age assurance process works prior to commencing the age assurance process for the service.

15.294 When implementing age assurance, service providers should have regard to the ICO's Children's code, and the Opinion.

Draft guidance

15.295 We are also expecting to publish accompanying guidance to the recommendations on highly effective, including additional technical detail and examples, to assist services in implementing highly effective age assurance in accordance with Measures AA1 to AA6.

15.296 The draft version of this guidance is set out at Annex 10.

Provisional conclusion

15.297 We consider that our recommendations on highly effective age assurance are justified for the purposes of ensuring consistency across the Children's Access Assessment, the Protection of Children Codes of Practice, and the draft Part 5 Guidance, as discussed above at sub-section 'Consistency with Part 5'. Consistency is important for providing regulatory clarity to services as to our expectations of how highly effective age assurance should be implemented.

15.298 In addition, we consider that our recommendations on highly effective age assurance are proportionate for several reasons.

15.299 Firstly, the use of highly effective age assurance as part of their systems and processes to prevent children from accessing PPC is a requirement under the Act for services who do not prohibit PPC in their terms of service (see Measures AA1 and AA3), and for the reasons set out above, our provisional view is that it is proportionate to recommend in connection with PC for some services under our other proposed measures (see Measures AA2, AA4, AA5 and

AA6).[145] It is therefore important to provide sufficient clarity to services in scope of the requirements and/or our recommendations as to how they can implement age assurance in such a way that is highly effective at correctly determining whether a particular user is a child. We consider that the criteria of technical accuracy, robustness, reliability and fairness are the minimum conditions required for services to secure this.

15.300 Second, we considered it proportionate to recommend steps that services should take to fulfil the four criteria, to meet our obligation under Schedule 4 of the Act to ensure that measures described are sufficiently clear, and at a sufficiently detailed level, that providers understand what those measures entail in practice.[146] We have ensured that these steps still provide service providers with flexibility to determine how they implement age assurance. We consider that this flexibility should benefit all services in scope of the age assurance measures, as it allows them to future-proof their systems and respond to technological developments over time in a way that is most cost effective for them.

15.301 In addition, we think that the principles of accessibility, interoperability, and transparency are important for services to have regard to so as to ensure they do not unduly prevent adult users from accessing legal content. We consider that consideration of these principles, and the recommendation for services to have regard to the relevant ICO guidance, are the minimum expectations required for services to fulfil their duties to have regard to users' rights to freedom of expression and rights to privacy under section 22 of the Act.[147]

15.302 We provide additional detailed analysis of the costs of implementing highly effective age assurance in Annex 12.

Other options considered

15.303 In developing our proposed measures, we considered different approaches relating to the scope and substance of our recommendations. We outline those considerations below.

Alternatives to highly effective age assurance

15.304 Services that do not prohibit PPC are required to use highly effective age assurance to comply with the duty to prevent children from encountering PPC.[148] We therefore did not consider recommending alternative means for identifying age for these services (see Measures AA1 and AA3).

15.305 The Act provides more flexibility for other U2U children's safety duties, including the duty to protect children in age groups judged to be at risk of harm from PC from encountering PC.[149] Here, age assurance is provided as an example of how a service can comply.

---

[145] Section 12(4), 12(5) and 12(6) of the Act.
[146] Paragraph 2(b) of Schedule 4 to the Act.
[147] Section 22 of the Act sets out that all services have, when deciding on, and implementing safety measures and policies, a duty to have particular regard to the importance of protecting users' right to freedom of expression within the law (section 22(2)); and, to the importance of protecting users from a breach of any statutory provision or rule of law concerning privacy that is relevant to the user operation of a user-to-user service (including, but not limited to, any such provision or rule of law concerning the processing of personal data) (section 22(3)).
[148] Sections 12(4)-(5) of the Act.
[149] Sections 12(4)-(5) and Section 7 of the Act.

15.306 Due to this flexibility, we considered whether highly effective age assurance or alternative approaches to identifying age would be appropriate in different circumstances. This includes where making recommendations relating to PPC and also for PC. In doing so, we had regard to the principle that more effective kinds of age assurance should be used to deal with higher levels of risk of harm to children, as set out in Schedule 4 of the Act.[150]

15.307 The alternative approaches we considered included recommending were:

- Self-declaration, rather than age assurance; and,
- A lower level of age assurance than the one expected from implementing the criteria for highly effective age assurance.

15.308 In the context of our proposals, we considered whether self-declaration might be proportionate to recommend for Measures AA2, AA4 and AA6 as they relate to the duty to "protect" children from encountering PC, rather than to "prevent" them from encountering PPC. However, as set out at sub-section 'Current Practice', our evidence indicates that self-declaration alone provides a low degree of certainty about the age of users. This is because children can and do easily circumvent it by providing a false age or date of birth. In addition, the Act sets out clearly that self-declaration is not age assurance.[151] While age assurance is not a requirement of the Act (with the exception of services who do not prohibit PPC), it is our view that age assurance is a vital component for ensuring that safety measures targeted at children can work effectively for those users.[152] We therefore consider it is not appropriate to recommend self-declaration alone under any of our measures.

15.309 Similarly, we considered the possibility of recommending highly effective age assurance for measures related to PPC and a lower level of effectiveness of age assurance for measures related to PC. However, based on current evidence, we do not believe that it would be feasible to specify an alternative level of effectiveness that is clearly distinguishable from highly effective age assurance and that would still achieve a sufficient level of protection for children relative to the risk of harm.

15.310 We also considered whether to allow providers discretion in determining what alternative approach to age assurance (other than highly effective age assurance) would be proportionate for their particular context. However, we believe this may not provide sufficient clarity to providers and could lead to a material risk that providers deploy ineffective age assurance methods that do not sufficiently protect children. The criteria-based approach allows flexibility for services to choose which steps to take to meet the standard of highly effective age assurance in a way that is technically feasible, appropriate and proportionate to their services.

15.311 As age assurance technology continues to evolve and we gather more evidence, we may consider this issue again in future.

Minimum age restrictions

15.312 Many services currently state a minimum age at which children can use the service in their terms of service. For social media services, this is often set at age 13+ which corresponds to data protection requirements relating to the processing of children's personal data without

[150] Schedule 4 (12)(d) of the Act.
[151] Section 230(4) of the Act.
[152] Section 12(7) of the Act provides that age assurance is an example of a measure which may be taken or used for the purpose of compliance with a duty set out in sections 12(2) or (3).

parental consent.[153] As discussed above at sub-section 'Current practice', our research shows that many younger children are creating their own profiles on online services despite these minimum age restrictions currently in place.[154]

15.313 The Act does not state that services have a duty to specify minimum age requirements, nor does it require services to operate any particular processes to enforce any such minimum age requirements where they do choose to set a minimum age limit for their service. However, where services have minimum age requirements, the Act requires U2U services to include in their terms of service details about the operation of those measures and to apply those terms consistently. We have already reflected this requirement in Terms of Service Measure TS1 in Section 19. We can take enforcement action if services fail to comply with this requirement.[155]

15.314 In developing our proposed measures, we considered whether it would be appropriate and proportionate to recommend that services that state a minimum age in their terms of service should use effective measures to enforce that provision, for instance, highly effective age assurance. We determined that this would not be proportionate given we have limited independent evidence that age assurance technology can correctly distinguish between children in different age groups to a highly effective standard and, given this, there is a risk that this could have serious impact on children's ability to access services.

15.315 There are separate age assurance considerations under Article 8 of the UK GDPR. We are not commenting on those requirements here.

Children in different age groups

15.316 In this first iteration of our Children's Safety Codes we are focusing on proposals that will result in safer, more protected experiences for all children, which are defined in the Act as users under the age of 18. The Act also requires all children to be prevented from encountering PPC and expects children *"in age groups judged to be at risk of harm"* to be protected from other harmful content.

15.317 We recognise that age is a key factor that will affect children's expectations and experiences of being online and our research indicates that certain online behaviours vary by age and developmental stage. However, there is currently limited evidence on the specific impact of harms to children in different age groups For example, the severity of impacts faced by children within particular age groups when exposed to PC may vary quite significantly and some children will be more vulnerable than others, even in older age groups. This includes neurodivergent children and children whose other characteristics such as a child's gender, race and sexuality may impact the harm they experience from content (see Volume 3 for further detail.) Therefore, while there may be some unintended adverse impacts on some children who would be less severely affected if exposed to such content, this may not be the

[153] Under Article 8 of the UK GDPR, it is unlawful for internet society services (ISS) to process the personal data of a child under the age of 13 unless consent has been given or authorised by the holder of parental responsibility. The data controller of the ISS should make "reasonable efforts" to verify that consent has been given or authorised. Similar systems exist in the US, the Children's Online Privacy Protection Act 1998 (COPPA) applies to any operator of a Web site or online service directed to children, or any operator that has actual knowledge that it is collecting or maintaining personal information from children under 13. The FTC published the COPPA Rule (16 CFR Part 312.5) requiring operators to obtain verifiable parental consent prior to any collection, use, and/or disclosure of personal information from children under the age of 13.

[154] Ofcom, 2024. Children's Online User Ages.

[155] For more information about Ofcom's enforcement powers, see the Illegal Harm Consultation, Chapter 29

case for all children across a particular age group for whom this additional protection may provide significant benefits.

15.318 In addition, there is currently limited independent evidence on the capability of current age assurance methods to correctly distinguish between child users of different ages to a highly effective standard, without disproportionately affecting children's rights. The use case for age assurance until now has predominantly been to identify which users are children and which are adults, and technology has developed to solve this problem. As a result, the technology for identifying the precise age of users below the age of 18 is still developing. While age verification methods requiring a photo-ID document (e.g., a passport) could identify the precise age of a user below the age of 18 they may risk excluding children from access to services they could otherwise benefit from in the absence of this documentation, which some children may not have.[156]

15.319 Given these limitations, our proposals focus at this stage on establishing recommended protections for all children under the age of 18, rather tailoring those protections for children in different age groups.

15.320 However, services may choose to apply a differentiated approach to children in different age groups when it comes to protecting children from encountering PC and NDC. For example, a service may judge that it is appropriate to expose older children to certain content that may be harmful to younger children. In these cases, we would expect them to be able to demonstrate how they have made this judgement and the methodology that they have used to establish what children are in different age groups, as well as how they are ensuring that younger children are protected from being exposed to harmful content. For the reasons set out above, we do not consider self-declaration would be a sufficiently robust method for targeting protections to children in different age groups.

15.321 As part of this consultation, we want to understand if our proposals are likely to have a disproportionate impact on children in different age groups, especially in relation to PC and in relation to older children and their rights and freedom to access information, and how we might be able to build more flexibility to mitigate these negative impacts while ensuring they receive the right protections from harmful content.[27]

[156] Obtaining a passport for a child carries a time and monetary cost that could be prohibitive for some families, for instance. Details on these costs can be found at Gov.UK, Get a passport for your child. [accessed 10 April 2024].

16. Content Moderation U2U

Content moderation is when a service provider reviews content to decide whether it is permitted on its service and how it will action that content. It can be done automatically, by humans, or by a combination of the two.[157] For the purpose of complying with the children's safety duties in the Act, content moderation involves reviewing content to decide whether it is content harmful to children and actioning it appropriately, to prevent or protect children from encountering it. Implemented effectively, content moderation systems and processes allow providers to take swift, accurate and consistent action on harmful content. As such, content moderation plays a hugely important role in combatting online harms.

We are proposing measures for user-to-user services likely to be accessed by children. All services should have in place systems and processes to swiftly action content harmful to children. For services that are multi-risk (regardless of size) and large services, we are proposing additional measures ensuring that content policies are clear, moderation functions are well-resourced, content moderators receive adequate training and content reviews are appropriately prioritised. These will support greater effectiveness of content moderation systems and processes, reducing in turn the prevalence of harmful content and the risks of harm to children. Ofcom's Guidance on Content Harmful to Children (Section 8, Volume 3) supports these measures by describing a non-exhaustive list of examples of kinds of content that Ofcom considers to be, or considers not to be, primary priority content and priority content that is harmful to children.

Our proposals allow services flexibility to implement measures in a way that is cost-effective and proportionate to the circumstances of the service, as long as they remain effective in addressing harm. The proposals in this section are not prescriptive about the balance services should strike between human and automated review of content, however we would expect that services would comply with data protection law and refer to ICO guidance.[158]

We are aware that large services with a substantial amount of content may rely on automated content moderation tools, in conjunction with human moderators, to ensure that the moderation of content harmful to children is scalable and efficient. We are not including specific recommendations in this consultation on the use of automated tools to identify and review content. We are planning an additional consultation later this year on how automated detection tools can be used to mitigate the risk of content harmful to children and illegal content. These proposals will draw on our growing technical evidence base and will complement the existing content moderation measures set out in our draft Codes of Practice.

In this chapter we propose to adopt an approach consistent with that outlined in our previous 2023 Illegal Harms Consultation. The measures to protect children that we propose in this section should be considered separately, and in addition, to those outlined in the Illegal Harms Consultation. That is because there are differences in the duties underlying these measures that are unique to protecting children from harm. We are also proposing an additional measure that volunteer moderators should be provided with materials for their roles, as this will help providers of services likely to be accessed by children to meet their duties. We are proposing this measure for inclusion in our Illegal Content Codes, given evidence we have considered regarding the use of volunteer moderators to identify and take down illegal content.

[157] Encyclopedia of Big Data, 2017. Content Moderation. [accessed 2 August 2023].
[158] Further detail on relevant ICO guidance can be found in the Rights Assessment section of CM1.

The proposals in this section are applicable only to U2U services. For our proposed recommendations for Search moderation, see Section 17.

Our proposals

#	Proposed measure	Who should implement this[159]
CM1	**Content moderation systems and processes are designed to swiftly take action against content harmful to children**	All U2U services
CM2	Set **internal content policies**	All U2U services that are either (or both): • Large • Multi-risk for content harmful to children
CM3	Set **performance targets** for content moderation function	
CM4	Have and apply **policies on prioritisation** of content for review	
CM5	Ensure content moderation functions are **well-resourced**	
CM6	Ensure content moderation teams are **appropriately trained**	
CM7	**Volunteer moderators** should be provided with **materials for their roles**	All U2U services that use volunteer moderation and are either (or both): • Large • Multi-risk for content harmful to children

Consultation questions

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

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

[159] These proposed measures relate to providers of services likely to be accessed by children.

What does content moderation entail?

16.1 Content moderation systems and processes differ from service to service and can involve tasks performed by humans, automated tools or a combination of the two. Service providers generally employ them to address a wide variety of harms, including illegal content and legal content on their service that does not comply with their own content policies i.e. violative content.[160] Content policies tend to dictate how violative content will be moderated. These policies are typically set out for users in external policies in the terms of service and community guidelines.

16.2 The overall effect of having a content moderation process is to help keep users, in particular children, safe, support compliance with legal obligations, and to maintain a trusted environment for other actors, such as advertisers.[161]

What risks does ineffective content moderation pose to children?

16.3 As set out in Ofcom's Children Register of Risks (Volume 3, Section 7) research indicates that, in the absence of well-designed and resourced content moderation systems, children are more likely to encounter harmful content. Specifically, a lack of effective and consistently applied content moderation processes can lead to an increased prevalence of harmful content and therefore a greater risk of children encountering it.

16.4 Ineffective or poorly resourced content moderation appears to have serious impacts on user safety across a wide range of harms. There is evidence of online services' content moderation systems failing to tackle content that is harmful to children,[162] as well as evidence from services reporting on their content moderation practices, that shows an increase in user safety and a reduction in harmful content when investment is put into improving content moderation systems.[163]

[160] For our proposed recommendations for content moderation for illegal content, please see Ofcom, 2023: Consultation: Protecting people from illegal harms online.

[161] Trust & Safety Professional Association (Singh, S.), 2019. What Is Content Moderation? Everything in Moderation: An Analysis of How Internet Platforms are Using Artificial Intelligence to Moderate User-Generated Content. [accessed 2 August 2023].

[162] In this study the researchers explored Instagram, TikTok, and Pinterest with avatar accounts registered as being 15-years-of-age. Content was identified and scraped using hashtags that have been frequently used to post suicide and self-harm related material. While this is a single study and may not represent all children's experiences, it demonstrates that this type of content was available on the services at the time of the study. The Bright Initiative and Molly Rose Foundation, 2023. Preventable yet pervasive: The prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest [accessed 06 March 2024].

[163] Google, 2020. Information quality and content moderation. [accessed 3 August 2023]; Reddit, 2022. Transparency Report. [accessed 3 August 2023]; Google, no date. Featured policies: Violent extremism. [accessed 3 August 2023].

16.5 Time pressures on human moderators and poor resourcing of moderation can increase the risk of human error in moderation decisions.[164] Periods where there is no human moderator presence on services may increase the risk that content harmful to children is widely viewed or disseminated before being actioned by the service in line with their terms of service.[165]

Interaction with Illegal Harms

16.6 In our 2023 Illegal Harms Consultation, we proposed the following measures regarding content moderation for U2U services to be included in our draft Illegal Content Codes:

- **Measure 4A**: Content moderation systems or processes are designed to take down illegal content swiftly.
- **Measure 4B**: Internal content moderation policies are set having regard to the findings of risk assessment and any evidence of emerging harms on the service.
- **Measure 4C**: Performance targets are set for content moderation functions and services measure whether they are achieving them.
- **Measure 4D**: When prioritising what content to review, regard is had to the following factors: virality of content, potential severity of content and the likelihood that content is illegal.
- **Measure 4E**: Content moderation teams are resourced to meet performance targets and can ordinarily meet increases in demand for content moderation caused by external events.
- **Measure 4F**: Staff working in content moderation must receive training and materials to enable them to identify and take down illegal content.

16.7 We provisionally consider that measures 4B, 4C, 4D and 4E in the draft Illegal Content Codes to be proportionate for providers of a service likely to be accessed by children. Further, we provisionally consider that measures 4A and 4F in the draft Illegal Content Codes are also proportionate for providers of a service likely to be accessed by children, once these are framed as ensuring that services achieve the children's safety duties.[166]

16.8 As with the draft Illegal Content Codes, we considered different approaches for these measures regarding whether to specify: a) detail for how services should configure content moderation systems and processes, b) the outcomes systems and processes should achieve, or c) factors services should have regard to in designing these systems and processes. Our

---

[164] A report by Demos highlighted that human content moderators have to make decisions in minutes, often about content in a language or from a context they do not understand, making mistakes inevitable. Source: Demos (Krasodomski, Jones, A.), 2020. Everything in Moderation: Platforms, communities and users in a healthy online environment.[accessed 4 October 2023]; A report by CASM Technology and ISD found a major increase in the number of antisemitic posts coinciding with a reduction in content moderation staff at one social media service, saying the analysis demonstrates "the broader and longer-term impact that platforms de-prioritising content moderation can have on the spread of online hate." Source: CASM Technology and ISD, 2023. Antisemitism on Twitter Before and After Elon Musk's Acquisition. Note: On its methodology, the report comments there are 'inherent challenges in training language models on as nuanced a topic as antisemitism, but this architecture is evaluated to operate with an accuracy of 76%. [accessed 3 August 2023]. Similarly, in late 2022, the Anti-Defamation League (ADL), noted an increase in antisemitic content on the same service and a decrease in the moderation of antisemitic posts. Source: ADL, 2022. Extremists, Far Right Figures Exploit Recent Changes to Twitter. [accessed 3 August 2023];

[165] Ofcom, 2022. The Buffalo attack: Implications for online safety . [Accessed 29 February 2024].

[166] Notably, unlike our draft Illegal Content Codes, equivalent measures for 4A and 4F do not include proposals for services to take down content.

provisional view remains that option c) is the most proportionate approach as it raises standards whilst also allowing for flexibility, given that there is no 'one-size-fits-all' approach to content moderation across the sector.[167] We set out below our detailed assessments of the evidence and impact of these measures as they relate to duties for services likely to be accessed by children.

16.9 We are also proposing to include an additional measure into the draft Children's Safety Codes regarding the provision of materials to volunteer moderators. We are proposing an equivalent measure for inclusion in our Illegal Content Codes (Measure 4G), given evidence we have considered regarding the use of volunteer moderators to identify and take down illegal content.

Our proposals to protect children

16.10 The Act requires U2U services likely to be accessed by children to, where proportionate, take measures relating to a number of areas including content moderation, in order to fulfil their children's safety duties.[168] Services also have a duty to take appropriate action in response to complaints about harmful content, and to handle appeals about action taken against content or a user during the moderation process.[169]

16.11 In developing our proposals for how providers of services likely to be accessed by children can meet these duties, we consider that effectively implemented content moderation – determining whether content is harmful to children and actioning it in line with a service's terms of service – is key to helping reduce the risk of children encountering harmful content.

16.12 We propose seven measures for the moderation of content on U2U services likely to be accessed by children. We discuss our detailed rationale including which services we propose these measures apply to in the rest of this section.

16.13 Measure 1 sets out the minimum expectations for content moderation that is applicable to all U2U services, as required by the Act. The measure has been designed so as not to be prescriptive, and therefore allows services flexibility to adapt this measure to their specific characteristics.

- **Measure CM1:** Services should have in place content moderation systems and processes designed to swiftly take action against content that is harmful to children.

16.14 Measures 2-6 are a package of additional, complementary measures applicable to all U2U services that are multi-risk[170] for content harmful to children (regardless of size) and all large[171] low-risk U2U services. These proposed measures recommend that such services take additional appropriate steps with regard to content moderation policies, performance targets, prioritisation, resourcing and training.

---

[167] Gillespie, T. et al. (2020) 'Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates', Internet Policy Review, 9(4). [accessed 29 February 2024]. Center for Democracy & Technology (2021). Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems. Center for Democracy & Technology. [accessed 2 August 2023].

[168] Section 12(2), (3) and (8) of the Online Safety Act 2023

[169] We consider Reporting, Complaints and Appeals in Section 18.

[170] See Section 14 within this Volume for a definition of a multi-risk service.

[171] See Section 14 within this Volume for a definition of a large service.

- **Measure CM2:** Services should set internal content policies having regard to at least the findings of their risk assessment and any evidence of emerging harms on their service.
- **Measure CM3:** Services should set performance targets for their content moderation functions and measure whether they are achieving these.
- **Measure CM4:** Services should have and apply policies on prioritising content for review, having regard to at least the following factors: virality of content, potential severity of content, the likelihood that content is harmful to children, including whether it has been flagged by a trusted flagger.
- **Measure CM5:** Services should resource their content moderation functions to give effect to their internal content policies and performance targets, having regard to specified factors.
- **Measure CM6:** Services should ensure their content moderation teams are appropriately trained.

16.15 Measure 7 supports the effective use of volunteer moderation among services that are multi-risk for content harmful to children (regardless of size) and all large[172] low-risk U2U services.

- **Measure CM7:** Services that use volunteer moderation, should provide moderators with materials for their roles.

16.16 While we have not included specific recommendations on the use of automated technologies in the first iteration of the draft code, we encourage service providers that already deploy these technologies as part of their content moderation processes to continue to do so and encourage those that are not currently deploying automated technologies for content detection to invest in systems that will help detect this content in their services at scale.

Relationship between Terms of Service and moderating content harmful to children

16.17 The Act requires that U2U services likely to be accessed by children include provisions in their terms of service which specify how children will be prevented from encountering PPC and those in age groups judged to be at risk will be protected from encountering PC and NDC, and must apply those provisions consistently. [173][174]

16.18 This is relevant to how services might action content that its moderation systems and processes identify as harmful. For example, where a service prohibits pornography for all users, once it has identified content as pornography, the service should apply its terms of service and remove the content. Where a service does not prohibit pornography in its terms of service for all users, allows children to use the service but prevents them from accessing that content through the use of age assurance, the service should apply its terms of service

[172] See Framework for Codes at Section 14 within this Volume for a definition of a large service.
[173] Section 12(9) and (10) of the Online Safety Act 2023
[174] In this first iteration of our Children's Safety Codes we are focusing on proposals that will result in safer, more protected experiences for all children, which are defined in the Act as users under the age of 18, see Children of different ages at Section 13.

and employ access controls so that children cannot access that content or the relevant part of the service on which that content is accessible to adult users.[175]

16.19 In addition, if a provider takes or uses a measure designed to prevent access to the whole of the service of a part of the service by children under a certain age, they are subject to a duty to –

- Include provisions in the terms of service specifying details about the operation of the measure; and
- Apply those provisions consistently.[176]

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

Explanation of the measure

16.20 This proposed measure recommends that services likely to be accessed by children should have in place systems and processes to action content that is harmful to children where they become aware of its presence on the service and proposes actions that services can take to assist them in fulfilling the children's safety duties.[177]

16.21 The Act does not require services likely to be accessed by children to take down content that is harmful to children for all users. The Act requires providers to use proportionate systems and processes, which includes content moderation, to prevent all children from encountering PPC and to protect children in relevant age groups from encountering other harmful content on a service.[178]

16.22 Services could achieve this outcome via access controls[179] to stop children from accessing the content or part of the service or applying other content moderation measures to protect children as appropriate. Some services may choose not to allow one or more type of content that is harmful to children on their service at all under their terms of service (i.e. to prohibit it on the service for all users). In that case, the content should be taken down once identified, in line with the service's terms of service.

16.23 The Act allows for service providers to have different terms of service for UK users when compared to users elsewhere in the world. In practice, where the Act requires content to be actioned, this means actioned for UK child users (or UK adult users if relevant). Service providers may limit their content moderation approach to UK users only, if they have a way of identifying them; alternatively, they may apply the same approach to all users if they prefer, in order to ensure that children in the UK are protected in line with the requirements in the Act.

[175] 'Access controls' refers to mechanisms to determine which users can access online content or spaces.
[176] Section 12(11) of the Online Safety Act 2023
[177] The Online Safety Act requires services to use proportionate systems and processes designed to prevent children encountering PPC and protect children from encountering PC and NDC.
[178] Section 12(2), (3) and (8)(e) of the Online Safety Act 2023
[179] Section 15 (Age Assurance), 'Access controls: mechanisms to determine which users can access online content or spaces', see Measures AA3 and AA4.

16.24 The children's safety duties require action to be taken against specified categories of content defined in the Act (i.e. the relevant categories of PPC and PC, as well as NDC).[180] We consider there are two approaches that providers of services likely to be accessed by children may take to fulfil their duties to prevent or protect children from encountering this content:

- Service providers may set about making judgements as to whether individual pieces of content should be classified as content that is harmful to children (by reference to Ofcom's Guidance on Content Harmful to Children if they wish[181]), for the express purpose of complying with the children's safety duties; or
- If service providers are satisfied that their terms of service are cast broadly enough to necessarily cover PPC, PC and NDC content, and secure that appropriate action is taken when that content is identified, service providers may choose to apply those when moderating content to secure compliance with the children's safety duties.

16.25 In both circumstances, their judgement should be made on the basis of all relevant information that is reasonably available to the service provider.[182]

16.26 Under both approaches, service providers should swiftly action content that is harmful to children that they have identified, to comply with the children's safety duties.

16.27 We consider that content is actioned swiftly if it is actioned within a reasonable timescale to effectively mitigate the risk of harm to children and meet their children's safety duties.

16.28 When implementing this measure, services should evaluate the type of content moderation systems and processes that are appropriate for their service. The minimum requirement for services is to have a complaints-based system to identify content for moderation, which a small or low-risk service may have if it is sufficient to allow them to comply with their duties effectively. Such services should allow users and affected persons to easily report content that is harmful to children and should have a process to assess these complaints as they arise and take appropriate action once they have determined if the content is harmful to children.[183]

16.29 We would expect many services to adopt content moderation systems and processes that go beyond this. Below, we set out further measures that all U2U services that are multi-risk for content harmful to children (regardless of size) and all large low-risk U2U services should take in respect of their content moderation systems.

16.30 As explained, how a service moderates this content will depend on circumstances such as whether they prohibit PPC, PC and NDC on the service. This measure is not prescriptive as to whether services use wholly or mainly human or automated content moderation systems and processes.

---

[180] These are defined in sections 60 to 62 of the Online Safety Act 2023

[181] Ofcom's draft Guidance on Content Harmful to Children sets out examples of content and kinds of content that Ofcom considers to be, and not to be, PPC and PC, which will be a useful resource to providers in understanding how to make judgements on content that is harmful to children.

[182] See section 192(2) of the Act.

[183] As required under sections 20, 21, 31 and 32 of the Online Safety Act 2023. User reports and appeals are types of complaint. We use complaints to refer to all types of complaints, including user reports and appeals. User reports are a specific type of complaint about content, submitted through a reporting tool. For more information, see Section 18 (Reporting and complaints).

16.31 As explained in Section 19, services likely to be accessed by children have a duty to include provisions in their terms of service which specify how they are going to prevent children from encountering PPC and protect them from encountering PC and NDC. We consider, in principle, those terms should point to what content moderation process they are using to ensure they protect children on their service.

Appropriate actions to prevent children from encountering Primary Priority Content

16.32 Where the provider of a service likely to be accessed by children prohibits PPC in its terms of service for all users, and it becomes aware of content that it suspects is PPC, it should review the content to determine whether it is in breach of those terms, and, if it determines that it is, swiftly action the content in line with its terms of service.

16.33 Below we set out a non-exhaustive list of further content moderation actions, that may not be mutually exclusive, that a service that prohibits PPC can take to protect children from PPC

- **Suspended functionality** – services may restrict a user's access to functionalities. For example, a service may prevent a user from commenting, posting content or messaging other users.
- **Ban and suspend users** – services may ban or suspend users from accessing their service. Services may use a warning or strike system to determine when a user should be banned or suspended.

16.34 Where the provider of a service likely to be accessed by children does not prohibit all kinds of PPC in its terms of service, but its principal purpose is not the hosting or dissemination of PPC, then our separate Age Assurance Measure AA3 recommends that it should implement highly effective age assurance to prevent children from encountering identified PPC.[184] The service may still allow children to access the service, but it should then take appropriate action such as using filtering – so that each piece of content identified as PPC is only visible to users confirmed to be adults using highly effective age assurance – or ensuring that all identified PPC is present only on parts of the service where access is restricted to users confirmed to be adults using highly effective age assurance.

16.35 The content moderation systems and processes that the service chooses to put into place should be designed in a way that ensures that identified PPC is actioned swiftly.

Appropriate actions to protect children from Priority Content and Non-designated Content

16.36 Where a service likely to be accessed by children prohibits any kinds of PC and NDC in its terms of service for all users and it has identified content that it suspects is PC or NDC, it should consider whether the content is in breach of those terms, and, if it is, action the content in line with its terms of service. For example, where a service prohibits content that encourages, promotes or provides instructions for a challenge or stunt highly likely to result in serious injury in its terms; and states that this content is prohibited for all users, the service should apply its terms of service and remove the content.

[184] On the other hand, where the principal purpose of a service is the dissemination or hosting of PPC, Measure AA1 recommends highly effective age assurance to stop children accessing the entire service. Where a service is not likely to be accessed by children, it is no longer in scope of the children's safety duties and these content moderation measures, as with other measures in the Codes, would not apply.

16.37 Below we set out a non-exhaustive list of further content moderation actions, that may not be mutually exclusive, that a service that prohibits PC and NDC can take to protect children from priority content.

- **Suspended functionality** – services may restrict a user's access to functionalities. For example, a service may prevent a user from commenting, posting content or messaging other users.
- **Ban and suspend users** – services may ban or suspend users from accessing their service. Services may use a warning or strike system to determine when a user should be banned or suspended.

16.38 Where the service provider does not prohibit one or more kinds of PC and NDC in its terms of service, and it has identified content that it suspects is PC or NDC, it should further moderate the content. If the content is determined to be PC that is not prohibited, the provider should swiftly action that content so as to protect children from encountering it. The action a service takes may depend on a number of factors, including the nature and severity of the harm and the age of the user, if known.

16.39 Below we set out a non-exhaustive list of content moderation actions, that may not be mutually exclusive, that a service that does not prohibit one or more kinds of PC can take, aside from content removal, to protect children from priority content. For example:[185]

- **Access and content controls** – Services may stop children from accessing pieces of content to protect them from encountering PC and NDC, for example by implementing filtering to prevent children from seeing certain violent content. Services may implement controls over parts of the service – such as communities, forums, groups or tabs – where certain PC or NDC appears, to stop children from accessing these.
- **Limiting the prominence of content** – Services may make content appear less frequently or prominently, for example in recommender feeds, on the service's home page or within search results provided by the service. This is sometimes referred to as downranking. We are recommending that services that have a recommender system that have identified a medium or high risk of at least one type of PC excluding bullying should reduce the prominence of content that is likely to be PC. For more information about our proposed Recommender Systems Measure RS2, refer to Section 20.
- **Overlays, interstitials, blurring and labels** - Services may accompany content with messages noting that the content may be harmful or sensitive. Services may offer links to supporting organisations or resources, including material within their sites. Overlays (or interstitials) and blurring may cover an entire piece of content and require the user to click through.[186] A label may provide a warning or additional context.

[185] Services have provided examples of content moderation actions that they can take. Ofcom, 2022. Ofcom's first year of video-sharing platform regulation [Accessed 22 February 2024]. Subsequent references are to this report throughout; Ofcom,2023. Content moderation in user-to-user online services [accessed 22 February 2024].

[186] Overlays or interstitials: Elements such as pop-ups, overlays or webpages which appear before the target content is displayed, or while navigating between pages. Typically, the user will need to take an action, such as clicking through, to reach the target content.

16.40 Certain services that do not prohibit at least one kind of PC are recommended to use highly effective age assurance to protect children from PC, as part of our separate Age Assurance measures. Please refer to Section 15 in this volume for more information.[187]

Effectiveness at addressing risks to children

16.41 Content moderation systems and processes are already employed by a number of services that are likely to be accessed by children. Annex 11 is a summary of child safety measures employed across platforms most used by children. All 33 platforms assessed stated that they employ a form of content moderation; with 22 of 33 further stating that they used AI to block types of harmful content or contact.[188] A number of services such as TikTok, Snapchat, Meta and YouTube have publicly spoken about both their human and automated content moderation systems and processes.[189] Our 2022 VSP report found that TikTok's content moderators manually review content when it reaches a certain level of popularity in terms of views and regularly undertake targeted searches of the platform for specific risks. On Snapchat the 'Spotlight' feed goes through human moderation before content can reach more than 25 viewers. The report also found that on Twitch all reports submitted by users are reviewed by trained specialists. These specialists also review content that is signaled by Twitch's automated tools and when these tools detect potentially harmful content, they are reviewed by a human before action is taken.[190]

16.42 As set out in Volume 3, Section 7.11, research indicates that, in the absence of well-designed and resourced content moderation systems, children are more likely to encounter harmful content. The proposed measures are designed to reduce the risk of children encountering harmful content.

16.43 Effectively implemented content moderation is a key way services can reduce the risk of children encountering harmful content by determining whether content is harmful to children and actioning it, in line with their terms of service. Conversely, a lack of effective and consistently applied content moderation processes can lead to an increased prevalence

---

[187] More specifically, Age Assurance measure AA4 applies to services that do not prohibit at least one kind of PC, do not have a principal purpose of hosting or disseminating PC, and have medium or high risk for at least one kind of PC they allow.

[188] Please refer to Annex 11 on Child Safety Measures research.

[189] Tiktok have on their website set out their approach to content moderation, they use both automated moderation technology to review videos uploaded and the use of human moderators to review content flagged by technology, reports from users, popular content and assessing appeals. TikTok. Our approach to content moderation [accessed: 14 December 2023]; SnapChat have on their website said they 'use a combination of automated tools and human review to moderate our public content surfaces (such as Spotlight, Public Stories, and Maps) – including machine learning tools and dedicated teams of real people – to review potentially inappropriate content in public posts'; Snapchat,2023. Snapchat Moderation Enforcement, and appeals [accessed: 14 December 2023]; Instagram have on their website said that they use artificial intelligence and human reviewers to moderate content that may go against its Community Guidelines. Instagram states that AI 'can detect and remove content that goes against our Community Guidelines before anyone reports it. At other times, our technology sends content to human review teams to take a closer look and make a decision on it'. How Instagram uses artificial intelligence to moderate content [accessed: 01 March 2024]; YouTube have on their website said 'machine-learning systems help us identify and remove spam automatically, as well as remove re-upload of content that we have already reviewed and determined violates our policies. YouTube takes action on other flagged videos after review by trained human reviewers.' YouTube Community Guidelines and policies - How YouTube Works [accessed 14 December 2023].

[190] Ofcom, 2022

of harmful content and therefore a greater risk of children encountering it (see Volume 3, Section 7.11).

Rights assessment

16.44 This proposed measure recommends where a service likely to be accessed by children has identified content that is harmful to children, it should review and action the content swiftly. Although there is no duty for platforms to remove such content (unlike illegal content, as discussed in our Illegal Harms Consultation), the Act requires providers to use proportionate systems and processes designed to prevent all children from encountering PPC and to protect children in age groups judged to be at risk of harm from PC and NDC.[191] As explained above, there are a variety of approaches that a service provider can take to secure the outcomes proposed by this content moderation measure. This may include taking down the content where it is prohibited under their terms of service, employing content controls[192] to prevent access to the content or part of the service on which it is allowed where it is permitted under their terms of service, as well as other actions such as banning or suspending users who are found to have shared this sort of content contrary to their terms of service, or limit the prominence of content, amongst others. In implementing this measure service providers may take content moderation steps which have a potentially significant impact on the rights of users (including both children and adults), in particular, their rights to privacy (Article 8), freedom of religion and belief (Article 9) and freedom of expression (Article 10) and freedom of association (Article 11). We have therefore considered the extent to which the degree of interference with these rights is proportionate.

16.45 By limiting children's exposure to content that is harmful to them in this way, the proposed measure will seek to secure adequate protections for children from harm, in line with the legitimate aims of the Act. The detection and moderation of content harmful to children acts to prevent the harmful consequences of such content that can be inflicted on them. These consequences can include harm to children's physical, mental or emotional wellbeing. We therefore consider that a significant public interest exists in measures which aim to prevent children from encountering PPC and protect them from PC and NDC. This substantial public interest relates to the protection of children's health and morals, public safety, and particularly the protection of the rights of others, namely child users of regulated services.

Freedom of expression and association

16.46 As explained in Volume 1, Section 2, Article 10 of the ECHR upholds the right to freedom of expression, encompassing the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. The right to freedom of expression is a qualified right. Ofcom must exercise its duties under the Act, considering users' and platforms' Article 10 rights, and not restrict that right unless it is satisfied that it is necessary and proportionate to do so.

16.47 With this proposed measure, potential interference with child users' freedom of expression arises where the service provider decides to apply content moderation processes to material it considers to be harmful to children as in this case the service provider would need, one way or another, to restrict children's access to it. In some cases (as noted below), this could

[191] Section 12(3) of the Act

[192] Content control mechanisms determine the visibility and accessibility of content including its removal or reduction. In this context, content controls include access controls such as blocking access to a part of the service that may host the harmful content. See also Section 15 and Measures AA3 and AA4.

also result in impacts on adult[193] users' ability to access the content as well. In addition, as a result of being found to have shared content that is harmful to children, some users might end up having their ability to use the service restricted in some way or removed (i.e. if their accounts were suspended or banned). This impact has a potential to be significant particularly if that judgement is incorrect (as in this case, there would not be a substantial public interest in access to the piece of content in question/their account being restricted).

16.48 However, the duty for services to treat content harmful to children appropriately is a requirement of the Act, and not of this proposed measure. The proposed measure does not involve services taking any steps in relation to content of which they are not aware, or to restrict access to any content which they do not judge to the harmful to children, and does not, in itself, require any particular actions to be taken against users who are found to have shared such content. This measure is designed in a way that is not prescriptive about how such content is to be moderated, instead it seeks to secure that the provider's systems or processes are designed so that they take steps to prevent all children from encountering (PPC) and protect children in age groups judged to be at risk of harm from encountering content harmful to them (PC/NDC) swiftly, where they become aware of its presence on the service. To the extent that the actions taken as a result of this measure affect children's ability to access or share content that is harmful to them and adult users' ability to share such content with children, we consider that is justified in line with the duties of the Act, as the benefits of the protections on children should outweigh the restrictions on other users' rights to encounter (if they are other child users) or share (whether they are children or adults) this form of content with children. Under the proposed measure, we would expect services to have in mind the duty to prevent children from encountering PPC compared to the duty to protect children from encountering PC and NDC, in determining what appropriate action to take.

16.49 We also consider that, while there is a potential risk for a margin of error in content moderation, services have incentives to limit the amount of content that is wrongly actioned, to meet their users' expectations and to avoid the costs of dealing with appeals. Where a service decides to take down content or restrict access to it on the basis that it is content harmful to children, complaints procedures required under section 21(2) of the Act should allow for the user to complain and for appropriate action to be taken in response. The complaints process may also mitigate the impact on the user's right to freedom of expression by giving the user a mechanism for redress and providing a route to rectify any negative impact by having their content restored to an equivalent position to the one it would have benefited from prior to the action being taken.

16.50 Impacts on freedom of expression could in principle arise in relation to the most highly protected forms of speech, such as religious expression (which could also affect users' rights to religion or belief under Article 9 ECHR) or political expression, and in relation to kinds of content that the Act seeks to protect, such as content of democratic importance and journalistic content.[194] However, we consider there is unlikely to be a systematic effect on

[193] Users also include those who are operating on behalf of a business, or accounts that might also be concerned with other entities, such as charities, as well as those with their own, individual account. Both corporate and individual users can benefit from the right to freedom of expression, and we acknowledge the potential risk of interference with the rights of these users to freedom of expression, in addition to the rights of children and adults as individuals. For ease of reference, when we refer to rights of adult users, we include those who are acting on behalf of a business or other entity.

[194] See the duties set out in sections 19 and 17 of the Act.

these kinds of content: for instance, such content would be unlikely to be particularly vulnerable to being wrongly classified as content harmful to children. In addition, we have provided examples of types of content, including more protected forms of speech, which we consider to be or not to be PPC or PC in Ofcom's draft Guidance on content harmful to children (Volume 3, Section 8), and we encourage service providers to have regard to this Guidance in implementing this measure.

16.51 While this is not a requirement of the measure, we acknowledge that a greater degree of interference with users' rights (both children's and adults' rights) could arise if the service provider chose to adopt terms of service which defined the content in relation to which children's access should be restricted more widely than is necessary to comply with the Act, or chose to prohibit one or more forms of PPC, PC or NDC for all users. In this case, services could also be restricting children's or adult users' access to certain types of content which is not required under the duties in the Act, and might also not be harmful, or might be less severely harmful, to them. However, it remains open to services as a commercial matter (and in the exercise of their own right to freedom of expression) to decide what forms of content to allow or not to allow on their service so long as they comply with the Act. Services have incentives to meet their users' expectations in this regard, too.

16.52 In addition to the impacts above, we have considered if there could be a risk of a more general 'chilling effect' if UK users (including both adults and children) were, as a result of this proposed measure, to cease to use regulated services which have implemented a more effective content moderation process. However, we do not consider that any such effect would be likely to arise both for the reasons set out above, and given that many UK users already use service providers which have implemented content moderation processes across their services.

16.53 The use of content moderation to limit children's exposure to harmful content could also have positive impacts on freedom of expression and freedom of association rights of children, for example, more effective moderation of content that is harmful to children could result in safer spaces online where children may feel more able to join online communities and receive and impart (non-harmful) ideas and information with other users. This measure could therefore also have significant benefits to children, in terms of safeguarding their rights to freedom of expression and assembly in safer online spaces, as well as in terms of protecting them from exposure to harm.

16.54 We therefore consider that the impact of the proposed measure as a result of services' content moderation decisions and processes on child and adult users' rights to freedom of expression, above and beyond the requirements of the Act, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that the proposed measure is therefore proportionate.

16.55 The proposed measure may also have an impact on services' rights to freedom of expression as, to the extent that they do not already choose to prohibit or limit children's exposure to the relevant forms of content that is harmful to children, services would need to put in place steps to ensure that it is appropriately dealt with in line with the measure. However, most of this impact arises from the duties placed on services under the Act by the UK Parliament, and we are allowing flexibility for services as to the precise approach and action they take to secure the outcomes required by the duties. We therefore consider that to the extent that the proposed measure impacts on services' rights to freedom of expression, it is likely to

constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

Privacy

16.56 As explained in Volume 1, Section 2, Article 8 of the ECHR confers the right to respect for individuals' private and family life. An interference with the right to privacy must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest. Again, in order to be 'necessary', the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.

16.57 All content moderation, whether by automated tools or human moderators, will involve the processing of personal data of individuals, including children. It will therefore impact on users' rights to privacy and their rights under data protection law. The degree of interference will depend to a degree on the extent to which the nature of their affected content and communications is public or private, or, in other words, gives rise to a legitimate expectation of privacy. This proposed measure is not limited only to content or communications that are communicated publicly,[195] and may lead to the review of content or communications in relation to which individuals might expect a reasonable degree of privacy, which would in turn lead to more significant privacy impacts than in connection with impacts on content and communications that are widely publicly available (whether on the service concerned or more generally). The impact on users' rights would also be affected by the nature of the action taken as a result of the content moderation process. For example, the level of intrusion and significance of the impact is likely to be higher where content is judged to be a form of content that is harmful to children or to violate the terms of service that would relate to the children's safety duties, and therefore would lead to a form of appropriate action, or where more restrictive measures are applied as a result compared to less restrictive measures.[196]

16.58 The duty for services to treat content harmful to children appropriately, including through the application of content moderation systems and processes, is a requirement of the Act, and not of this proposed measure, and we are giving services flexibility as to precisely how they implement this and what action they take. We recognise that depending on how service providers decide to implement the proposed measure, it could result in a greater or lesser impact on users' privacy rights. However, as noted above, it remains open to services (and in the exercise of their own rights to freedom of expression) to decide what forms of content to allow or not to allow on their service, and what forms of personal data they consider they need to gather to enforce their content polices, so long as they comply with

[195] As part of its consultation on illegal harms Ofcom consulted on draft guidance on content communicated 'publicly' and 'privately' under the Online Safety Act. That guidance recognises that whether content is communicated 'publicly' or 'privately' for the purposes of the Act will not necessarily align with whether that content engages users' (or other individuals') rights to privacy under Article 8 of the European Convention on Human Rights. For example, it is possible that users might have a right to privacy under Article 8 of the ECHR in relation to content which is communicated 'publicly' for the purposes of the Act. Conversely, users may not have a right to privacy under Article 8 of the ECHR in relation to content which is nevertheless communicated 'privately' for the purposes of the Act.

[196] We have, for example, also proposed to recommend that certain services should use highly effective age assurance to ensure children are prevented from encountering PPC or PC identified on the service (see section 15 (Age assurance) Measure AA3 and Measure AA4), which might have potentially significant privacy impacts. However, we have separately assessed the impact of those proposed measures and we do not repeat this discussion here.

the Act and the requirements of data protection legislation. Under the proposed measure, we would expect service providers to have in mind the duty to prevent children from encountering PPC compared to the duty to protect them from encountering PC and NDC in determining what appropriate action to take. To the extent that this would require additional restrictions on users' privacy rights or processing of additional personal data, we consider this proportionate as a result.

16.59 We acknowledge the potential risk of negative impacts on the right to privacy, for example where content is categorised as harmful to children incorrectly, or where there is a disproportionate level of monitoring of content (automated or manual) by providers where the risk of content harmful to children is low. The use of content moderation is one of a number of measures we have recommended as ways for providers to comply with their duties under the Act. We do not anticipate that providers will rely solely on this to reduce the likelihood of children encountering or being harmed by content harmful to them. We expect that services will make use of all appropriate measures to assist them in complying with their duties, which may include measures that are less intrusive from a privacy perspective and carry less risk of an impact on users' privacy rights. We have assessed the impact on rights on the basis that services and providers will utilise the full suite of relevant measures set out in the draft Children's Safety Code, as required to achieve compliance with the child safety duties in the Act.

16.60 The degree of impact will also depend on the extent of personal data about individuals which may need to be processed to give effect to the applicable content moderation processes. The proposed measure does not specify that service providers should obtain or retain any specific types of personal data about individual users as part of their content moderation processes, and we consider that service providers can implement the measure in a way which minimises the amount of personal data which may be processed or retained so that it is no more than needed to give effect to their content moderation processes. In processing users' personal data for the purposes of this measure, services would need to comply with relevant data protection legislation. This means they should apply appropriate safeguards to protect the rights of both children, whose personal data may require special consideration,[197] and adults. Providers may also use third parties to carry out content moderation on their behalf and ICO guidance is clear that services should ensure that individuals' rights to privacy are fully protected when a third party has access to their personal data. [198]

16.61 Insofar as services use automated processing in content moderation (which we are not specifically recommending), services should refer to ICO guidance on content moderation to determine whether the processing is solely automated i.e. has no meaningful human involvement, and results in decisions that have a legal or similarly significant effect on users.[199] of their content moderation processes, services should comply with the standards set out in the ICO Children's Code in respect of children's personal data, along with other relevant guidance from the ICO.[200] [201]

---

[197] In line with Recital 38 UK GDPR

[198] Further information on the requirements for contracts between data controllers and processors can be found at Contracts and liabilities between controllers and processors

[199] In which case Article 22 UK GDPR requirements are likely to apply

[200] ICO, Children's code guidance and resources and other relevant such as Online safety and data protection

[201] Such as Online safety and data protection

16.62 We therefore consider that (assuming service providers also comply with data protection legislation requirements) the impact of the proposed measure as a result of services' content moderation decisions and processes on child and adult users' rights to privacy, above and beyond the requirements of the Act, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

Impacts on services

16.63 In order to implement this measure, a provider of a service likely to be accessed by children would incur direct costs associated with putting in place new systems and processes, or adjusting existing ones to moderate content harmful to children. We expect that the costs of doing this will vary by service. We set out below the potential foreseeable impacts across services.

16.64 Smaller, low risk services are unlikely to receive a high number of complaints regarding content harmful to children. As such, services will only have a limited amount of content to review. They are unlikely to require a complex content moderation system to review such content effectively and appropriately action content to give effect to their terms of service. This type of service might therefore decide to implement a simple system, with complaints assessed sequentially, to meet the minimum requirement of the Act. Doing so may entail some small, one-off costs of designing and implementing such a system. Ongoing costs associated with moderators reviewing the content and actioning where appropriate are likely to vary in proportion to the size and risk level of a service and therefore are expected to be small for small low-risk services.

16.65 Larger and riskier services will typically face higher costs to develop content moderation systems and processes in line with the children's safety duties. The costs of implementing this measure are likely to include both one-off costs of developing a system and ongoing costs of maintaining it. In terms of one-off costs, for services that decide to build their own systems internally, these costs may include hiring experienced content moderation systems designers, developing content moderation tools, project management and integration with data analytics/measurement software. For services which are not building their systems internally, the main cost would be the adoption of third-party moderation solutions and integration with their internal policies, tools and processes as well as the fees that they pay the third party to moderate for them. There would also be several ongoing costs relating to systems maintenance, hosting and data logging which would vary by service, as well as the ongoing costs related to detecting, reviewing, and actioning content harmful to children. We consider that the costs discussed here reflect the base level of cost which is required to design and operate a content moderation function to action content harmful to children and consider that a proportionate approach for large and riskier services will also entail costs additional to this, as set out in the measures CM2 to CM7 below.

16.66 The costs associated with taking appropriate action in relation to content may depend on whether a service prohibits all kinds of content harmful to children. If it does, then any such content would be expected to be removed from the service altogether once identified. However, where a service does not prohibit certain kinds of content harmful to children, it may take different actions to protect children, which can entail higher cost. Examples include implementing filtering or blurring so that certain pieces of content are not visible for some users, or restricting access to some parts of the service (such as tabs, forums or

communities where content harmful to children appears). The relevant actions and associated costs are highly dependent on the context of a specific service, its architecture and any existing systems used for access or content control. Therefore, we cannot quantify these, but we expect that costs would typically be higher where a service has a relatively complex architecture, which is more likely on larger services.

16.67 Note that, where services are recommended to implement highly effective age assurance for the purposes of applying access or content controls, the cost of this is assessed separately in Section 15, and is not part of our assessment for this measure. The same applies for costs associated with implementing our Recommender System measures, where we assess impacts separately in Section 20.

16.68 Services may incur costs related to ensuring that their content moderation functions are treating different kinds of content in a way that is aligned to the definitions of different kinds of content harmful to children set out in the Act. Services may consult our guidance on content harmful to children, which aims to provide clarity to services and can assist them in making sure they identify and action relevant kinds of content appropriately (see Volume 3, Section 8).

16.69 Overall, we expect that the costs of implementing this measure will vary widely between services. For the smallest low-risk services, costs are likely to be negligible or in the small thousands at most. For some large or risky services these costs could extend to multiple millions depending upon the approach taken, the volume of content on the service and/or the volume of reports received.

16.70 We note that service providers likely to be accessed by children will also be in scope of the related measure proposed in our Illegal Harms Consultation and consider that there is likely to be substantial overlap between the two measures in terms of set-up costs, as it may be practical for services to use common systems and processes in relation to both illegal content and content harmful to children.[202] This includes small, low-risk services, which could easily adapt a simple, complaints-based system designed to deal with illegal content to cover content harmful to children alongside illegal harms.

16.71 We also note that many services will already have in place content moderation systems to review and action content that they consider to be harmful to children, which would reduce the incremental cost associated with this proposed measure. However, the type of content identified and actioned under existing systems may not be the same as that set out in the Act, and therefore some reconsideration or reconfiguration of content moderation systems may be needed. Such changes may also lead to a higher volume of content being flagged for review than previously, increasing content moderation costs.

16.72 While the costs described above may be significant for some services, these have not been fully quantified because we believe this measure captures the minimum steps to ensure a basic level of content moderation that would be proportionate for all U2U services to comply with the children's safety duties. We consider that a proportionate approach for large or riskier services will also entail costs additional to this, as set out in the subsequent measures.

[202] Please see paragraph 12.47 of Ofcom's Illegal harms consultation

Which providers we propose should implement this measure

16.73 As discussed above, we consider that the measure outlined here captures fundamental steps for content moderation that represent the minimum expected from providers of any service likely to be accessed by children to meet the children's safety duties. Evidence shows that having content moderation systems and processes in place is necessary for services to comply with the children's safety duties. Such systems and processes allow services to identify harmful content on the service and to take appropriate action to keep children safe, reducing children's exposure to such content. We therefore propose that this measure should apply to all U2U services likely to be accessed by children.

16.74 We believe that the impact on services is mitigated by the flexibility of this measure, as we are not being prescriptive as to how services implement content moderation systems and processes, allowing services to take cost-effective processes that are proportionate to the context of each service. We expect costs to scale with both service size and risk. We expect that small services that are not multi-risk for content harmful to children can appropriately action content using simpler, less costly systems and processes, and the moderation costs to a small service that receives very few or no user reports are expected to be minimal.

16.75 We therefore consider that this measure is proportionate for all U2U services likely to be accessed by children.

16.76 However, for services that are large or multi-risk for content harmful to children, we consider that this measure alone would be insufficient. As such services operate in a more complex risk environment, we consider it proportionate to further specify how they should design their policies, processes, frameworks and resources to moderate content effectively. The proposed measures CM2-CM7 discussed in the rest of this section consist of a package of further steps that we recommend services should take if they are large or multi-risk for content harmful to children.

16.77 For smaller services which are not multi-risk for content harmful to children, we expect this measure CM1 to provide adequate protection. Such services are less likely to face high volumes of diverse content that is potentially harmful to children that they need to assess. Given that these services operate in a simpler risk environment, they could reasonably be expected to meet their child safety duties without employing more sophisticated formal processes and frameworks implied by measures CM2-CM7. These services are likely to have relatively limited resources and we consider that the benefit to children's safety may be greatest if the services have flexibility to focus resources on core systems and processes for identifying and actioning any harmful content, rather than diverting resources towards additional, more complex systems and processes that may have only small incremental benefits on such services. In any case, smaller services that are not multi-risk should still take all necessary steps to give effect to Measure CM1.

Provisional conclusion

16.78 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm U2U services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU B1 in Annex A7.

Measure CM2: Set internal content moderation policies

Explanation of the measure

16.79 Content policies often exist in two forms - external and internal - and explain how harmful content should be identified and actioned:

a) External content policies are set out in publicly available documents aimed at users of the service, providing an overview of a service's rules about what content is and is not prohibited. These normally form part of a service's publicly facing terms of service and have names such as 'community guidelines', 'community policies', and 'community standards'. Users are expected to understand and observe these rules when posting content on services.
b) Internal content policies are usually more detailed versions of external content policies which set out rules, standards or guidelines, including around what content is and is not prohibited.

16.80 Further, internal content moderation policies provide a framework for how policies should be operationalised and enforced. These policies are used as a guide for enforcement by content moderators and other relevant teams, as well as designers of automated systems to assist in identifying potential content breaches.

16.81 We propose that all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level should set clear internal content moderation policies and keep a written record of these policies. When setting internal policies, services should have regard to at least the findings of their risk assessments and have processes in place to update these policies in response to any evidence of emerging harms on their service. In addition, services may refer to Ofcom's Guidance on Content Harmful to Children.[203]

16.82 We consider that there would be significant benefits in recommending that services have regard to at least their risk assessments and have processes in place to update these policies in response to evidence of emerging harms when setting their internal policies. Both of these data sources would provide evidence about the challenges a service's content moderation functions face. It is reasonable to infer that such data would enable services to make higher quality decisions about what to put in their internal content moderation policies. This should improve the quality of these policies and by extension improve the performance of services' content moderation systems, thereby reducing harm to children.

Effectiveness at addressing risks to children.

16.83 Industry stakeholders suggests that setting internal content policies may play a key role in establishing an effective content moderation system, particularly for U2U services multi-risk for content harmful to children. For example, several large and medium providers publicly

[203] This represents a slight clarification of the wording we proposed for the equivalent measure in our Illegal Harms Consultation. We would use the same wording in both.

state that external and internal content moderation policies play a key part in keeping users safe online.[204]

16.84 Services such as Instagram, YouTube, and Discord have published their external content moderation policies which provide an overview of their prohibited and not prohibited content on the service.[205] Further, service providers such as TikTok and the Mid-Sized Platform Group have also spoken on the need for content moderation policies.[206]

16.85 We understand that internal policies may go beyond the scope of external policies, they may be far more detailed, with more definitions, exceptions and examples, and that these policies, unlike external policies, may remain unpublished.[207] It is our understanding that publishing information on internal policies may be used by users to circumvent the content moderation systems and processes.[208]

16.86 Responses to the 2023 Protection of Children Call for Evidence from civil society organisation such as the Samaritans and Carnegie UK Trust recommended that services establish and enforce comprehensive internal content moderation policies.[209]

16.87 Services that are multi-risk for content harmful to children (regardless of size) and all large low-risk U2U services may have large volumes of diverse content to moderate. Putting in place clear internal content moderation policies and keeping a written record of these will ensure consistency, accuracy and timeliness of decision making.

16.88 Therefore, internal content policies have a number of potential benefits, notably: increasing efficiency, increasing accuracy and consistency of decision-making which should reduce the amount of time harmful content to children remains accessible to children on a service and setting out how content moderation decisions need to take user rights into account. We understand that there may be trade-offs with accuracy and speed, but having internal content policies will increase consistency in decision making and ensure harmful content is appropriately removed.

---

[204] TikTok, 2019. Creating Policies for Tomorrow's Content Platforms | TikTok Newsroom [accessed 1 February 2024]; YouTube, 2019. The Four Rs of Responsibility, Part 1: Removing harmful content. Twitter, no date. The Twitter Rules.

[205] Following Molly Russell's death, in 2019 Instagram changed their policy regarding graphic and non-graphic self-harm related content, Instagram Policy Changes on Self-Harm Related Content | Instagram Blog [accessed 14 December 2023]; Meta, 2024, Suicide and Self-Injury | Transparency Centre [accessed 11 March 2024]; YouTube on their website have published their Child Safety Policy, including harms such as 'harmful or dangerous acts involving minors' and cyberbullying, Child safety policy - YouTube Help (google.com) [accessed 14 December 2023]; Discord in October 2023, published a 'suicide and self-harm policy explainer', Discord,2023 Suicide and Self-Harm Policy Explainer | Discord [accessed 29 February 2024].

[206] TikTok in 2020 said that platforms should 'look to approach the protection and safety of their users through policies, product, people, and partners'. TikTok, 2020. Creating Policies for Tomorrow's Content Platforms | TikTok Newsroom [accessed 14 December 2023]; The Mid-Sized Platform Group which comprises Patreon, Eventbrite, Reddit, Pinterest, Vimeo and TripAdvisor, said they all have strong commitments to their users to keep them safe, identify malicious actors on their platforms and create a positive online experience, which are exemplified by each of their content policies. Source: Mid-Sized Platform Group, 2022. [accessed 14 December 2023] .

[207] Khoury College at Northeastern University, no date. Content Moderation Techniques. [accessed 3 August 2023].

[208] Alan Turing Institute, 2021. Understanding online hate: VSP Regulation and the broader context. pg. 90[accessed 3 August 2023];

[209] Samaritans' response to 2023 Protection of Children Call for Evidence.; Samaritans,2023. Online Harms guidelines [accessed 14 December 2023]; Carnegie response to 2023 Protection of Children Call for Evidence.

Rights assessment

16.89 This proposal recommends services in scope of the measure have in place internal content moderation policies that take account of risks identified in their risk assessment. This option is designed flexibly in a way that does not tell services how to moderate content harmful to children, just that there are internal content policies outlining how to moderate it.

Freedom of expression and association

16.90 We consider that this proposed measure has the potential to impact on users' (both adults' and children's) rights to freedom of expression for the reasons set out in relation to Measure CM1 above.

16.91 In addition to the impacts identified in Measure CM1, we are of the view that this measure has the potential to interfere with users' rights to freedom of expression if internal content moderation policies defined the content in scope of these policies more widely than is necessary to comply with the Act. Although it is open to services to make a commercial decision about the type of content they allow on their service, in this proposed measure we are recommending that services take account of the findings of their risk assessments and any evidence of emerging harms, to ensure any interference with these rights are kept to a minimum and are proportionate in relation to the risk of harm to children. Internal content moderation policies can set out a level of detail that may not be practical to do in external facing policies, providing content moderators with greater clarity on the type of content that is harmful to children, resulting in a higher degree of content being flagged as content harmful to children.

16.92 We consider there may also be positive impacts on users' (including adults and children) right to freedom of expression and freedom of association from services implementing this proposed measure. Where services are likely to be dealing with large volumes of content, the process of considering these matters in advance and preparing a policy would tend to improve internal scrutiny, and improve the consistency and predictability of decisions, in a way which we think would also tend to protect users' rights.

16.93 It should result in fewer instances of content being incorrectly identified as content harmful to children, allowing individuals to express their views and receive or impart information that is not content harmful to children, without unjustified interference. In doing so, online spaces would be made safer for children, affording them greater opportunities to exercise their right to freedom of association.

16.94 We therefore consider that the impact of the proposed measure recommending services set internal content moderation policies, on child and adult users' rights to freedom of expression and freedom of association, above and beyond the requirements of the Act, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

Privacy

16.95 We consider that this proposed measure has the potential to impact on users' (both adults' and children's) right to privacy for the reasons set out in relation to Measure CM1 above.

16.96 In addition to the impacts identified in Measure CM1, to the extent that, in setting internal content policies, services describe or define the content they are dealing with under the children's safety duties in a way which involves reference to information in respect of which

a user would have a reasonable expectation of privacy, or to personal data, users' rights in relation to these would be engaged. Services are required to comply with data protection laws[210] and internal policies should be drafted in a way that supports compliance. Having a set of policies in place would also encourage consistency to content moderation decisions, which will be a benefit to users as the nature of content harmful to children can change over time and as new functionalities are developed.

16.97 Where services are likely to be dealing with large volumes of content, the process of considering these matters in advance and preparing a policy would be likely to improve internal scrutiny, and improve the consistency and predictability of decisions, in a way which we think would also be likely to protect users' privacy and personal information rights.

16.98 We therefore consider that (assuming service providers comply with data protection legislation requirements) the impact of the proposed measure recommending that are large or multi-risk for content harmful to children, set internal content moderation policies, on child and adult users' rights to privacy, above and beyond the requirements of the Act, to be relatively limited, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

Impacts on services

16.99 In order for a service provider which does not already have an internal content policy to implement the measure, it would incur the full costs of developing such a policy. For a smaller U2U service, we anticipate that developing such a policy could take a small number of weeks of full-time work and involve legal and regulatory staff, and online safety/harms experts. For example, based upon our wage estimate assumptions as set out in Annex 12 if a service required 3 weeks of time across professional occupations (legal/regulatory staff) and 4 hours of senior leadership time to develop an internal content policy, this would represent a cost of approximately £3,000 to £7,000.

16.100 In some cases, services may use external experts which could increase costs. Engagement and approving new policies may also take up senior management's time, which would add to the upfront costs.

16.101 However, larger services may require more complex content policies, as the way in which harm can materialise is likely to be more varied on such services and the governance requirements needed to implement them are also likely to be more complex. Some services may use external experts which could increase costs. Engagement and approving new policies may also take up senior management's time, which would add to the upfront costs.

16.102 These factors may increase costs, due to the increased amount of time required to design more complex policies. These costs could reach the tens of thousands or more. In addition, there may be some small ongoing costs to all U2U services to ensure these policies remain up to date over time e.g. to take into account emerging harms.

16.103 Some providers of services likely to be accessed by children will also be in scope of the related measure proposed in our Illegal Harms Consultation.[211] We consider there may be

[210] Including the ICO's Children's Code and any relevant guidance from the ICO

[211] See our Illegal Harms Consultation, Volume 4, Chapter 12, page 35 for a full explanation of the measure.

some overlap between the measures, for example where similar guidelines may apply relating to how certain aspects of the policies are operationalised and enforced. However, any such overlaps and associated cost synergies may be limited, given the very different natures of the two types of harms.

16.104 Likewise, some services will already have policies in place which at least partly address this proposed measure. For these services, the proposed measure may mainly involve costs to update existing policies in line with risk assessments and any emerging evidence of harms.

16.105 We believe that these costs are mitigated by the flexibility of the measure, as we are not being prescriptive as to what should be included in an internal content policy, but instead propose to set out high-level requirements that give services flexibility to decide how to achieve what is proposed. This flexibility will allow them to take an approach proportionate to the risks they carry. A small service with medium risk of two kinds of content may choose to have a simpler internal content moderation policy, whereas a large and complex service that has identified high risk in relation to many kinds of content would be expected to develop a more complex and detailed policy.

16.106 We also consider there may be some countervailing benefits to services, as having these policies in place should enable staff to carry out content moderation more efficiently.

Which providers we propose should implement this measure

16.107 All U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) pose multiple significant risks of harm to children, and we therefore consider that the benefits of applying this measure to them are likely to be material. Our analysis suggests that for such services, the presence of internal content policies is an important part of an effective content moderation system which helps reduce this risk of harm. As outlined above, the absence of effective content moderation significantly increases the risk of harmful content being accessible to children. The costs of this measure are likely to scale with the number and level of risks and so will scale with benefits. We therefore consider that it would be proportionate to apply the measure to all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size).

16.108 The benefits of extending this proposed measure to large services that are not multi-risk for content harmful to children will be smaller, as the scope to reduce harm will be more limited where risk of harm to children is more limited. However, we still consider that having internal content moderation policies in place for such services will have important benefits for users. We have taken into account that the nature and prevalence of content which is harmful to children can change over time, meaning that even if a large service is currently low-risk, this could change over a short period of time (e.g. due to unforeseen changes in their user base or the type of content which is present on their service). Having an internal content moderation policy in place will help ensure that, if there were to be an increased risk of harm to children on such services, this would be dealt with quickly, reducing the resulting harms, which on a large service would have the potential to affect a lot of users, including children. The policy may also promote consistency in approach where a service has many moderators, which may be the case on a large service even if low-risk. We also note that large services are likely to have sufficient resources to develop or adjust these policies in line with the proposed measure. We therefore consider that it would be proportionate to apply this measure to all large U2U services (regardless of risk level).

16.109 As explained previously in relation to Measure CM1, at this stage we are not proposing to recommend this measure for smaller services which are not multi-risk for content harmful to children. We consider that the benefits of internal content moderation policies are likely to be materially smaller for services which are neither large nor multi-risk. They are less likely to face a diverse range of content that is potentially harmful to children that they need to assess. Therefore, we consider that the benefits of having a formal, structured framework in an internal content policy would be more limited, and that these services should be able to protect children by focusing resources on the implementation of Measure CM1.

16.110 We are therefore proposing that this measure should apply to all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level).

Provisional conclusion

16.111 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm U2U services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU B2 in Annex A7.

Measure CM3: Set performance targets for content moderation systems related to speed and accuracy

Explanation of the measure

16.112 We propose that all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level) should set performance targets for their content moderation functions and track whether they are meeting these targets.

16.113 Performance outcomes, usually in the form of KPIs, provide a quantitative measure of the effectiveness and efficacy of content moderation efforts. They help evaluate the performance of content moderation systems and processes by tracking specific metrics and comparing them against predefined targets or benchmarks.

16.114 We do not propose to stipulate the performance targets that services should set. However, we would propose in this measure, that at a minimum these should include targets relating to the time that a service takes to review or action content harmful to children and targets relating to the accuracy of content moderation decisions, for instance, implementing a quality assurance process. When setting targets, services should balance the desirability of swiftly actioning content against the desirability of making accurate moderation decisions.

Effectiveness at addressing risks to children

16.115 Some services record a wide range of metrics in relation to content moderation systems and processes. While many services record the same or similar metrics, there is considerable variation in precise definitions and naming conventions. The Trust & Safety Professional Association (TSPA) draws together these various metrics into five broad categories:

enforcement volume metrics;[212] time-based metrics;[213] quality metrics;[214] appeals metrics;[215] and other metrics.[216]

16.116 We understand that many services already set performance targets for the operation of their content moderation functions and measure whether they are achieving these, particularly related to speed. For example, TikTok records its removal rate within 24 hours[217], and Snapchat records 'Turnaround Time'[218] and publishes the median time for various platform violations.[219] Vimeo told us it aims to review and make determination on 80% of flagged content within 24 hours.[220] [**CONFIDENTIAL✂**].[221]

16.117 We also understand that in recent years some service providers have introduced metrics reflecting the viewing of violative content, before the content was actioned, which some providers see as particularly important – Meta described them as "the number we hold ourselves accountable to"[222] and YouTube "the primary metric [we use] to measure our responsibility work".[223] Further, some services also track the rate of appeals as a measure of the accuracy of the decisions that are taken.[224]

16.118 We consider that setting performance targets and measuring whether these are being achieved is likely to deliver important benefits. Where services are clear about the content moderation outcomes they are trying to achieve and measure whether they are achieving them, they will be better able to plan how to configure their systems and processes to meet

---

[212] 'Enforcement Volume Metrics' represent counting events that are part of the moderation process, such as capturing the volume of content flagged for review, the volume of content closed by a service's content moderation system, and the number of instances where a moderation action was taken. Trust & Safety Professional Association, no date. Metrics for Content Moderation. [accessed 3 August 2023].
[213] 'Time Based Metrics' are based on the amount of time taken to perform various parts of the content moderation process, such as review time, response time, removal time and time to action, i.e. the time between content being uploaded or created and a completed decision about whether the content is violating. Trust & Safety Professional Association, no date. Metrics for Content Moderation. [accessed 3 August 2023].
[214] 'Quality Metrics' are generally based on re-checks of previous reviews by either the existing review teams, subject matter experts, or dedicated quality reviewers. Trust & Safety Professional Association, no date. Metrics for Content Moderation. [accessed 3 August 2023].
[215] 'Appeals Metrics' involve re-checks of previous reviews by either the existing review teams, subject matter experts, or dedicated quality reviewers based on appeals, such as overturns and overturn rate, successful appeal rate, and time to resolution. Trust & Safety Professional Association, no date. Metrics for Content Moderation. [accessed 3 August 2023].
[216] 'Other Metrics' tend to be less directly tied to day-to-day operational decisions, such as prevalence, cost and impressions. Trust & Safety Professional Association, no date. Metrics for Content Moderation. [accessed 3 August 2023].
[217] TikTok,2023. Community Guidelines Enforcement Report.
[218] Snapchat Turnaround Time: The duration of time between when Snapchat's Trust & Safety teams first start to review a report (usually when a report is submitted) to the last enforcement action timestamp. If multiple rounds of review occur, the final time is clocked at the last action taken.
[219] Snapchat,2023. Transparency Report Glossary [accessed 11 March 2024].
[220] Vimeo response dated 8 July 2022 to the VSP Year 1 information request dated 6 June 2022.
[221] [**CONFIDENTIAL✂**].
[222] Violative content refers to content that breaches a service's terms of service, we understand that these may align or go beyond the harms included in our duties. Facebook (2018), 'Understanding the Facebook Community Standards Enforcement Report. [accessed 04 March 2024]
[223] YouTube's (2021), 'Building greater transparency and accountability with the Violative View Rate'. [accessed 04 March 2024]
[224] Pinterest,2023. Digital Services Act Transparency Report | Pinterest Policy. [accessed 04 March 2023]; Twitch,2022. H2 2022 Transparency Report (twitch.tv) [accessed 04 March 2023].

these goals and better able to optimise the operation of these systems. By configuring these systems and processes based on clear outcomes the service wishes to achieve, there will be a reduction in reviewer or system bias which could potentially leave children unprotected. We consider there are particular benefits from capturing both speed and accuracy targets, while allowing flexibility for services to determine how best to balance these in the context of their specific policies and procedures. On the other hand, an exclusive focus on speed could lead to poor decisions that either leave harmful content available to children or over-removes content that should be available; an exclusive focus on accuracy could result in services in content remaining available for a long time.

Rights assessment

16.119 This proposal recommends that services in scope should set performance targets as set out above, as part of their internal content policies recommended by Measure CM2 above. This proposed measure should therefore be seen as part of a package of measures relating to content moderation for content harmful to children, including Measures CM1 and CM2, for which we have assessed the rights impacts above. This option is designed flexibly and is not designed in a way that specifies the targets that services should meet, instead services should consider targets that measure the time it takes to action content moderation and the degree of accuracy of their content moderation systems.

Freedom of expression and association

16.120 We acknowledge the risk that setting performance targets can lead to a focus on speed rather than accuracy, which could result in incorrect content moderation decisions that could infringe on users' rights to freedom of expression. However, our proposal includes the recommendation that services also set performance targets for accuracy, which should mean that both speed and accuracy are considered by services, resulting in greater transparency and reliability in content moderation systems. This, we believe, would have a positive effect on users' rights to freedom of expression, by using content moderation systems that are implemented efficiently and accurately to identify content harmful to children. The benefits to children would be that online spaces are made safer for children by reducing the likelihood and period that content harmful to children is present on the service. This could positively impact children's rights to freedom of expression and freedom of association as children would be able to more safely engage with communities and content online. There could also be positive impacts on adult users' rights as content can be shared appropriately with the result that freedom of expression is preserved. We therefore consider that any interference to users' rights to freedom of expression arising from this proposed measure would be relatively limited and proportionate.

Privacy

16.121 We consider that measures to encourage an increase in accuracy of content moderation can have a positive impact on individuals' right to privacy. This proposed measure is intended to reduce the frequency of incorrect content moderation decisions, which can interfere with the right to privacy, particularly if inaccurate personal data is processed by the service and not rectified, for example where a user has been sanctioned due to an incorrect decision around content. It should also result in greater transparency and accountability around decisions that are made in relation to content moderation and the personal data that is inevitably processed in these actions.

16.122 However, we also acknowledge the risk that setting performance targets can lead to a focus on speed rather than accuracy, which could interfere with users' right to privacy. We have designed this measure so that services will need to balance the speed of decisions made with the degree of accuracy, which we think will mitigate the risk of unjustifiable interference with users' rights.

16.123 We therefore consider that any interference to users' rights to privacy arising from this proposed measure would be relatively limited and proportionate.

Impacts on services

16.124 Service providers are expected to incur direct costs if they would need to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

16.125 For a service provider which does not currently have performance metrics and targets in place to implement the measure, it would incur both one-off costs in designing metrics and setting them up, including relevant data management processes, plus ongoing costs to track actual performance against targets. The flexibility given to services regarding how to implement this measure means that costs are likely to vary widely between services.

16.126 For a smaller service, we expect that the costs of designing or selecting a small and relatively simple set of metrics would be limited. The bulk of one-off costs for such services may include creating and implementing the relevant processes to track the time between when content is reported and when it is assessed and/or action is taken. A simple bespoke system designed to capture this and also estimate accuracy – based solely on the outcome of user appeals – could take around a month's design, development, testing and implementation. Based on our cost assumptions set out in Annex 12, if this required around 30 days of software engineering time, this could represent a cost of around £8,000 to £16,000.[225] Alternatively, a service might opt to licence a third-party ticketing system at a relatively low cost – such solutions are available from around £50/month per staff user.

16.127 For large services, or those with medium or high risk for many kinds of content harmful to children, the number and complexity of metrics themselves, and of associated data management processes, may be significantly higher and entail higher costs. For such services it may also be proportionate to design and automate systems for proactive QA of moderation decisions, which would introduce complexity. Therefore, one-off costs could run from the tens to hundreds of thousands depending on the service design and volume of reports, which is likely to be linked to service size and number of risks.

16.128 As well as implementation costs, there would also be ongoing costs including to measure performance against these metrics (e.g. analytics teams), and analysing these metrics, as well as data storage costs. To assess the accuracy of content moderation decisions, services are likely to need to take a sample of those decisions and re-assess them. We have not quantified these costs as they are likely to vary greatly depending on the characteristics of a service. For example, a small service with relatively limited user-to-user functionality (e.g. text-based comments only), low volumes of content, and medium risk for two kinds of content harmful to children only, may be able to track performance against a single or small number of simple accuracy targets, using a simple and targeted QA process. On the other hand, costs may be very material where services have larger and more diverse types of

[225] Assuming 30 days FTE software engineer time.

content which pose material risk across many kinds of content harmful to children, potentially requiring more complex and extensive metrics relating to accuracy and speed of actioning content, and greater resource to conduct QA across a large sample of decisions.

16.129 For providers of services likely to be accessed by children that are also in scope of the related measure proposed in our Illegal Harms Consultation (i.e. services which are large or multi-risk in relation to Illegal Harms), we consider that there may be some overlaps between the two measures due to similarities in the nature of the proposals.[226] The types of metrics and the systems or processes used to track against targets are likely to be similar. Therefore, we expect that the one-off costs associated with the proposed measure will be lower for services that are also in scope of the related Illegal Harms measure. In terms of ongoing monitoring of performance against these metrics, there may be substantial cost overlaps to the extent that such monitoring is automated, but less so where it is more reliant on human input.

16.130 Some services, particularly larger ones, may already have processes or metrics in place which at least partly address this proposed measure. For these services, the proposed measure may involve any costs of adjusting existing approaches to ensure the recommendations of the proposed measure are met.

Which providers we propose should implement this measure

16.131 For services with a large volume of content to assess, we consider that there would be important benefits from setting performance targets for their content moderation functions and tracking whether they are met. As set out above, we consider that services that follow this measure are more likely to operate effective content moderation systems, mitigating the risk of harm to users.

16.132 Although the costs of this measure are significant, we consider that the benefits are likely to be sufficiently important to justify this proposal for all large services (regardless of risk level) as well as smaller services that are multi-risk for content harmful to children, given the fundamental role that effective content moderation plays in protecting users from harm. Large low-risk services may still have significant volumes of cases for moderation, and this measure should help to ensure that, if there were to be an increased risk of harm to children on such services, this would be dealt with quickly and accurately, reducing the resulting harms, which on a large service would have the potential to affect a lot of users, including children. Also, we do not propose to be prescriptive on the details of the performance targets set or how they are achieved, leaving scope for services to tailor these targets according to the risks that they identify and the specific operation of their services. This means that, for smaller services with fewer medium or high risks, where the benefits of the measure may be lower, we also expect costs to be lower.

16.133 As explained previously in relation to Measure CM1, we are not proposing at this stage to recommend this measure for smaller services which are not multi-risk for content harmful to children. We consider that implementing Measure CM1 would involve such services having regard to the speed and accuracy of their decisions, but that such services would benefit from greater flexibility in doing so. We consider that the specific approach to performance tracking proposed in this Measure CM3 would not be proportionate for these services, as they are likely to face lower volumes of potentially harmful content to moderate. Such

[226] See our Illegal Harms Consultation, Volume 4, Chapter 12, page 38 for a full explanation of the measure.

services may have more limited resources and we consider that the benefit to children's safety may be greater if they focus resources on the core systems and processes for identifying and actioning any harmful content, rather than necessarily investing in additional processes to track performance. We believe that Measure CM1 would provide adequate protection on such services.

16.134 We are therefore proposing that this measure should apply to all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level).

Provisional conclusion

16.135 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm U2U services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU B3 in Annex A7.

Measure CM4: Have and apply policies on prioritising content for review having regard to several factors

Explanation of the measure

16.136 We propose that all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level) should have and apply policies on prioritising content for review. We consider that where a service provider adopts a prioritisation framework, it is likely to result in higher quality decisions about what content to prioritise for review. Logically, we would expect this to result in a material reduction in harm to children, thereby delivering significant benefits.

16.137 We consider providers should take at least the following prioritisation criteria into consideration:

- The virality of a piece of content that is harmful to children and encountered by children,
- Potential severity of content, including whether it is likely to relate to content harmful to children. For example, in some circumstances PPC may have a higher severity than PC or NDC. Additionally, there may also be varying degrees of severity within these harms; and
- Likelihood that content is PPC, PC or NDC which could be based on signals available to the service that suggests the content is likely to fall into one of the harms categories e.g. when content is flagged by a trusted flagger.

Effectiveness at addressing risks to children

16.138 Due to their substantial user base, large U2U services often have significant volumes of content flagged to them as potentially harmful to children. Smaller services with material risks in relation to multiple types of content harmful to children are also likely to have different types of potentially harmful content to moderate at once. Providers of both types of services face difficult decisions about what content to prioritise for review. The decisions they take regarding prioritisation can have a material impact on harm caused to children. For example, if a provider chooses to review multiple pieces of PC which have not been viewed

by many (or any) children, before it reviews a piece of viral PPC that has been viewed by large numbers of children, this decision could result in significant harm to children.

16.139 Many providers use systems and processes to help them prioritise content for review. Providers with large-scale content moderation do not typically review content in chronological order but consider a range of factors, including: the virality of the content, its severity and the context of it becoming known to the provider (for example, whether or not as a consequence of a user report or other complaint). For example, TikTok says it recently started refining their approach to better prioritise accuracy, minimise views of violative content, and remove egregious content quickly.[227]

16.140 Facebook prioritises content that is expected to attract significant viewing.[228] Additionally, Facebook prioritises items based on how confident an algorithm is that moderators will agree that the content is violative and also on the 'severity' or 'egregiousness' of a suspected violation – arguably linked to the degree of harmfulness.[229] However, one side effect of this is that relatively less popular, less harmful items may remain available for long periods of time.[230]

16.141 Prioritising content relies on providers making trade-offs between a number of important goals, including harm, users' freedom of expression, and user experience. Trade-offs of this type may be unavoidable in a context of finite moderation capacity.[231] We currently think that providers are best placed to make these decisions based on their individual needs.

16.142 We consider that where a service provider adopts a prioritisation framework which considers the factors listed above (as well as other factors they identify as relevant), it is likely to result in higher quality decisions about what content to prioritise for review, as opposed to reviewing complaints in a chronological order. The benefits of having such a framework would likely be smaller for services which are neither large nor face material risks. This is because they are likely to receive materially fewer complaints for review, though they are still required to act promptly, they will have less of a need to prioritise between the complaints they do receive.

16.143 We explain below why we consider each of the prioritisation criteria covered by our proposed measure are important and relevant:

The virality of content harmful to children

16.144 "Virality" is a term used to describe the degree to which online content spreads easily and/or quickly across many online users, alongside how much engagement and/or views a piece of content received (i.e. 'shares', 'likes', and 'view', etc.).

---

[227] TikTok says it has upgraded the systems that route content for review, to better incorporate a video's expected reach (based on an account's following) when determining whether to remove it, escalate for human review, or take a different course of action. TikTok, 2023. Evolving our approach to content enforcement. [accessed 4 March 2024].

[228] Ofcom, 2023. Content moderation in user-to-user online services: An overview of processes and challenges, p.19-20.

[229] Violation refers to content that is prohibited by a service in its terms of service.

[230] Ofcom, 2023, Content moderation in user-to-user online services.

[231] Facebook said it uses a combination of technology and human review to prioritise content for moderation and review. When determining which content human review teams should review first, it considers how likely is it could lead to harm, both online and offline, how quickly it is being shared, and the likelihood it violates the platform's policies. Meta, 2022. How Meta prioritises content for review. [accessed 4 March 2024].

16.145 If a piece of harmful content is viral it has the potential to cause harm to larger audiences and may increase the likelihood of children encountering that harm. Prioritising the review of viral content means service providers can minimise the impact of the harm more efficiently.

16.146 In recognition of the impact of content virality and reach on user safety, some providers commonly use algorithms to ensure that reviewers' in-trays are prioritised in such a way as to minimise overall harm (e.g. by prioritising content that is most likely to be most harmful and/or to be viewed by the largest number of people).

16.147 We know that several of the larger services consider 'virality' of content when prioritising content for review, including both the 'likely' virality and 'actual' virality.[232]

16.148 However, we note that it is important to balance virality alongside other factors, including those listed here, as prioritising virality alone may mean other harms are missed. For example, "content promoting eating disorders, while not widely disseminated across the general user base, may circulate extensively within eating disorder communities and groups, posing a significant risk to those exposed to it."[233] We consider that some providers may take virality within segmented harm areas into account, for example they may choose to prioritise viral content within relevant policy areas such as eating disorders. We propose that virality should be considered alongside other metrics.

Severity of harm

16.149 We know that several providers already consider the severity of harm when prioritising content for review.[234] We acknowledge that our proposed recommendations regarding how providers should prioritise content harmful to children will need to be made alongside the prioritisation of illegal harms. In deciding how to prioritise the review of their content, providers should consider the impact that accessing harmful content is likely to have on children and the core objective of the Act to ensure that children are offered higher protections than other users in line with their specific vulnerabilities, regardless of whether it is illegal content, PPC, PC or NDC. We therefore propose that in considering how to prioritise content, providers should also take into account the likely severity of harm that might occur as a result of children's exposure to the content, given the nature of the content.

16.150 When considering the potential harm to children, providers also should have regard to the duties set out in the Act that all children should be prevented from encountering PPC, and children in age groups judged to be at risk of harm should be protected from encountering PC and NDC. In Ofcom's view, the fact that the Act is clear that the objective of the children's safety duties is to seek to prevent all children from encountering forms of PPC suggests that providers may consider generally giving a higher priority for review content that they have reasonable grounds to suspect may be PPC (compared to PC and NDC), as timely review of such content is more likely to result in swift actioning of this content, and therefore achieve the objective of preventing children from encountering it.

16.151 In some circumstances, PPC may have a higher degree of severity compared to PC and NDC. In the Children's Register of Risks, we set out the impacts of children encountering PPC

---

[232] Meta, 2020. How We Review content. [accessed 04 March 2024]; TikTok, 2023. Evolving our approach to content enforcement. [accessed 28 March 2024].
[233] BEAT response to 2023 Protection of Children Call for Evidence.
[234] Ofcom, 2023. Content moderation in user-to-user online services.

which can be severe, and in some cases, fatal.[235] Even within certain harms, there may be degrees of severity that need to be considered. For example, in its report into online hate, the Alan Turing Institute noted that "different types of online hate inflict different degrees and types of harm".[236] Some types of harmful content may have the capacity to result in more severe harm to children than others, such as those that have a degree of immediate direct harm compared to those that do not.[237] For example, the immediacy of livestreamed PPC, such as suicide or self-harm content, may require real-time moderation, or moderation that is faster than non-livestreamed content, so it may be appropriate to prioritise these.

16.152 We propose that providers should have regard to both the Illegal Harms and Children's Registers of Risk as well as findings from their risk assessments, and what they indicate about severity of harm, when considering their policy on prioritisation decisions. This is so when providers come to review content they can carefully consider factors around severity of harm, alongside other factors such as the freedom of expression implications that arise from reducing the visibility and spread of content, the context of the content, etc.

The likelihood that content is PPC, PC or NDC including whether it has been flagged by a trusted flagger

16.153 All else being equal, prioritising content for review where the signals available to the service suggest that there is a high likelihood that it is PPC should increase the speed with which content harmful to children is actioned, thereby making it more likely that children are prevented from encountering it and reducing harm to children. Similarly, it would be relevant for services to consider signals available which suggest that there is a high likelihood that content may be PC or NDC, when considering prioritisation of content for review.

16.154 User reports and complaints are likely to be the first way in which some services may find out about content harmful to children, particularly for those services which are not making extensive use of proactive detection methodologies. Complaints are already commonly used to help prioritise content for review, and they can potentially flag content harmful to children that other content moderation functions may have missed. For example, Twitch prioritise user reports based on the classification of the report and the severity of the reported behaviour.[238] However, we recognise that users may report content for various reasons and not all of these reasons are specifically for content that is harmful to children. This means that content being reported may not be a perfect indicator of breaches of services' content policies (which can require nuanced and context specific assessments). Therefore content flagged in this way may not be the most reliable source to enable providers to accurately identify content likely to be PPC, PC or NDC for prioritisation.

16.155 Trusted flaggers are any entity for which the provider has established a separate process for the purposes of reporting content which may include content harmful to children, based on the entity's expertise. For example, this could include individuals, NGOs, mental health organisations and other entities that have demonstrated accuracy and reliability in flagging content. Trusted Flaggers, often equipped with specialised knowledge and expertise, can provide valuable insights into the nature of the content and its potential harm. These signals

[235] See Volume 3, Section 7, Children's Register of Risks
[236] Ofcom and The Alan Turing Institute, 2021. Understanding online hate: VSP Regulation and the broader context. [accessed 25 August 2023].
[237] Meta, 2020. How We Review Content. [accessed 4 March 2024].
[238] Twitch, 2023. H1 2023 NetzDG Transparency Report. [accessed 4 March 2024].

from Trusted Flaggers can be particularly crucial in identifying and addressing content that is harmful to children but may not manifest as highly viral in the broader online community. For example, as Trusted Flaggers are particularly effective at identifying harmful content that violate its community guidelines, YouTube prioritise content reported by Trusted Flaggers because their flags have a higher action rate than the average user.[239]

16.156 Dedicated reporting channels (DRCs), used by Trusted Flaggers and Internet Referral Units, are sometimes used by services to flag potentially harmful content for review.[240] Though we are not currently recommending the use of DRCs and Trusted Flaggers in this iteration of the Code, where services currently employ them we propose that they should give priority for review to content flagged via these channels.[241] This is because where services have DRCs in place, the fact that a complaint comes from a Trusted Flagger or another expert body is of obvious relevance in determining what priority to give it as, all other things being equal, children may be unaware of harmful content and such complaints from Trusted Flaggers are more likely to be accurate and to reflect the Trusted Flagger's assessment of harm. They therefore have significant potential to reduce harm to users by alerting services to content that may be harmful to children that might not otherwise come to their attention (or might not otherwise be prioritised as swiftly).

Rights assessment

16.157 Our proposed measure recommends that services in scope have, as part of their internal content policies recommended by Measure CM2 above, policies on prioritising content for review, taking into account various factors such as the virality of that content, severity of harm and the likelihood that content is content harmful to children. This proposed measure should therefore be seen as part of a package of measures relating to content moderation for content harmful to children, including Measures CM1 and CM2, for which we have assessed the rights impacts above.

Freedom of expression and association

16.158 We do not consider that setting and applying a content prioritisation policy would in itself have any specific adverse impacts on users' or services' rights to freedom of expression or association. Instead, we think that this proposed measure would likely have a positive impact on the right to freedom of expression as it is aimed at services taking action on content harmful to children in a more targeted manner than would otherwise be the case if they had no prioritisation criteria to consider. We think having policies that make clear which content will be prioritised for review, including by focusing on the harm that may arise, should help secure that the highest risk content is actioned most swiftly. This should assist in securing services moderate content in a way that safeguards against disproportionate impacts on users' rights to freedom of expression. If the result is that users, particularly children, are better protected from harm, it may also have a positive impact on

[239] Google, 2020. Information quality & content moderation. [accessed 4 March 2024].
[240] Internet Referral Units are government-established entities responsible for flagging content to internet platforms that violate the platform's terms of service. Examples include the EU Internet Referral Unit (EU IRU) and the UK's Counter Terrorism Internet Referral Unit (CTIRU). [accessed 4 April].
[241] While some services currently use Trusted Flaggers for some illegal content, we do not currently have sufficient evidence on the effectiveness or cost of these programmes to recommend their use more generally for content harmful to children. For full consideration, please see Section 18 Reporting and Complaints.

children's freedom of expression and association as they may feel safer in using such services.

Privacy

16.159 We do not consider that setting and applying a prioritisation policy would have any additional impacts on users' privacy rights beyond those already considered in connection with Measures CM1 and CM2 above.

Impacts on services

16.160 Service providers are expected to incur direct costs if they would need to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

16.161 Services which do not currently have a prioritisation framework would incur one-off costs in designing and setting this up. We expect these would be largely one-off costs involving a small number of weeks of full-time work and involve legal, regulatory, ICT staff as well as online safety/harms experts, while agreeing the policy would likely need input from senior management. For example, if designing and setting up a relatively simple prioritisation framework (such as a smaller service with just two risks and a more limited quantity of content to review) required around three weeks FTE from professional occupations (legal, regulatory, ICT) and one day from senior leadership, this would be equivalent to costs of £4,000 to £7,000 using our salary assumptions as set out in Annex 12. However, for a larger and more complex service with a greater number of risks and a multitude of different metrics that can indicate virality, severity and suspected type of content, costs could be substantially higher than this, potentially reaching tens of thousands or more.

16.162 Services may incur costs related to assessing the prioritisation criteria, such as systems for determining whether content is likely to be harmful. While these activities may be costly depending on the approach taken, for example where they rely on machine learning models, we do not quantify these costs as we do not recommend any specific steps, leaving flexibility for providers to consider the appropriate approach for their services.

16.163 Depending on the approach taken to prioritisation, there may be ongoing costs with applying the policy, which can depend on whether this is a mainly manual or automated process.

16.164 Whilst automated prioritisation processes may be more cost-effective for services with large volumes of moderation cases, we consider that simpler processes – including manual ones – may be workable for some smaller services without adding large costs to the moderation process.

16.165 There are also likely to be some smaller ongoing costs for all services in ensuring that the prioritisation policy remains reflected in system design, and in reviewing it when appropriate. These costs are mitigated by the proposed measure not specifying exactly how services should prioritise content, giving services some flexibility in what they do.

16.166 For providers of services in scope of this measure who are also in scope of the related measure proposed in our Illegal Harms Consultation, we consider that there may be some overlaps between the two measures and that the estimated direct costs to these services of implementing this proposed measure would be reduced as a result.[242] For example, metrics

[242] See our Illegal Harms Consultation, Volume 4, Chapter 12, page 41 for a full explanation of the measure.

related to virality are likely to be similar or the same for both illegal content and content harmful to children. These services would need to consider how they can extend or adapt their existing framework to cover how suspected content harmful to children is prioritised appropriately.

16.167 Moreover, we expect that there will be countervailing benefits to services from implementing prioritisation, as it can enable content moderation functions to operate more efficiently.

Which providers we propose should implement this measure

16.168 We believe a prioritisation framework can contribute materially to the safety of children online, as it helps to ensure that services focus their content moderation resources on addressing pieces of content that are more likely to cause severe harm and to affect many children.

16.169 We consider that the benefits of adopting a prioritisation framework for service providers that are multi-risk for content harmful to children (regardless of size) are sufficiently important to mean it is proportionate for these providers to incur the costs of doing so, given the risk that they pose to children and the diversity of content types their moderation functions may be dealing with. As the proposed measure does not specify exactly how services should prioritise content, services have flexibility to shape their approach to be proportionate to the number and level of risks which are on their service.

16.170 The benefits of recommending this proposed measure to large services that are not multi-risk for content harmful to children will be smaller, as the scope to reduce harm will be more limited. However, similarly to other measures in this section, we still consider that having a prioritisation policy in place for such services will have important benefits for users. Even where a large service is currently low-risk, this could change over a short period of time (e.g. due to unforeseen changes in their user base or the type of content which is present on their service). Having a prioritisation policy in place will help ensure that services respond efficiently to such circumstances, reducing the resulting harms which, on a large service, would have the potential to affect a lot of users, including children. The policy may also promote consistency in approach where a service has many moderators, which may be the case on a large service even if low-risk. We also note that large services are likely to have sufficient resources to develop or adjust these policies in line with the proposed measure. We therefore consider that it would be proportionate to apply this measure to all large services.

16.171 As explained previously in relation to Measure CM1, at this stage we are not proposing to recommend this measure for smaller services that are not multi-risk for content harmful to children. We believe that this Measure CM4 would not be proportionate for these services as the benefits of having a prioritisation framework are likely to be materially lower, as they are likely to deal with a less diverse set of content moderation cases at scale. We expect that such services would benefit from greater flexibility in how they organise their content moderation function. They may have relatively limited resources and we consider that the benefit to children's safety may be greater if they focus resources on the core systems and processes for identifying and actioning any harmful content, rather than necessarily investing in additional frameworks. We believe that Measure CM1 would provide adequate protection on such services.

16.172 We are therefore proposing that this measure should apply to all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level).

Provisional conclusion

16.173 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm U2U services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU B4 in Annex A7.

Measure CM5: Ensure content moderation functions are well resourced

Explanation of the measure

16.174 We propose that all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level) should ensure that their content moderation functions are well-resourced so as to ensure that their internal content policies are fulfilled (including as to prioritisation)[243] and performance targets[244] are met. For such services, we consider that well-resourced content moderation systems – whether human, automated, or a combination of the two – and processes are key to effectively actioning content swiftly and mitigating the risk of children encountering harmful content.

16.175 At this stage, we do not think it would be beneficial for us to specify in detail how services should resource their content moderation functions. However, we do consider that there are factors to which services should have regard to when deciding how to resource their content moderation function, and that considering these is likely to result in important benefits. These factors are language expertise and resources and meeting spikes in demand for content moderation driven by external events. We explain below why these factors are important.

Effectiveness at addressing risks to children

16.176 Well-resourced content moderation functions enable services to review potential content harmful to children more quickly and make more accurate decisions as to whether to action it.

16.177 Ofcom's research suggests that, all other things being equal, a service may be able to reduce the 'turnaround time' between content being uploaded, reviewed and actioned by hiring more moderators, thereby reducing the amount of time that potentially harmful or violative content is 'live'.[245] In response to the 2023 Protection of Children Call for evidence, a number of stakeholders including Carnegie UK, Common Sense Media, EVAWG, Glitch, and the Molly Rose Foundation noted the importance of services investing sufficiently in human

[243] See Measures 2 and 4, above.
[244] See Measure 3, above.
[245] Ofcom, 2023. Content moderation in user-to-user online services. Page 26.

moderation resources to tackle harmful content.[246] In Ofcom's research into children's experiences of suicide, self-harm and eating disorder content, young people suggested implementing more human moderation rather than relying on what was perceived to be ineffective artificial intelligence as a way of improving safety features.[247]

Language expertise and resources

16.178 Given the large number of languages that are spoken in the UK and the fact that some services may target specific communities of language speakers, content posted in many languages has the potential to cause harm to users in the UK. We therefore propose that services consider the language capabilities that may be required to review potentially harmful content which could affect children in the UK and resource their systems accordingly.

16.179 We are aware that several services already consider the language content is posted in and/or ensure they have the language expertise within their moderation systems to deal with it, using both humans and automated methods to do so.[248] For example, Facebook and Instagram state that its content review team is global and reviews content 24/7 in over 70 languages.[249] TikTok also moderate content in more than 70 languages globally and provide information about the primary languages their moderators work in globally.[250]

16.180 In their response to the 2023 Protection of Children Call for Evidence, Common Sense Media noted that decisions from moderators 'require language fluency, cultural nuance, context of speech'.[251] Glitch noted that services should 'ensure moderation considers local context, including (but not limited to) linguistic, social, cultural, historical, racial and gendered context'.[252] Resolver (formerly Crisp) told us in stakeholder engagement that moderating in the English language is not enough when it comes to ensure child safety in the context of suicide and self-harm content and that it is necessary to understand culture and colloquialisms as well.[253]

16.181 The language expertise required to deal with the risk of harm in a particular language will likely differ from service to service based on a number of factors, including user base, content type and functionality. For this reason, we propose our Codes of Practice should not be prescriptive around what exact language expertise and resource is required on any

---

[246] Carnegie UK response to 2023 Protection of Children Call for Evidence; Common Sense Media response to 2023 Protection of Children Call for Evidence; EVAW response to 2023 Protection of Children Call for Evidence; Glitch response to 2023 Protection of Children Call for Evidence; Molly Rose Foundation response to 2023 Ofcom Protection of Children Call for Evidence.

[247] Ofcom, 2024. Experiences of children encountering online content promoting eating disorders, self-harm and suicide.

[248] "The social media companies said they moderated content or provided fact-checks in many languages: more than 70 languages for TikTok, and more than 60 for Meta, which owns Facebook. YouTube said it had more than 20,000 people reviewing and removing misinformation, including in languages such as Mandarin and Spanish; TikTok said it had thousands. The companies declined to say how many employees were doing work in languages other than English." Hsu, T., Misinformation Swirls in Non-English Languages Ahead of Midterms. The New York Times, 12 October. 2022. [accessed 3 August 2023].

[249] Facebook, 2023. DSA transparency report. [accessed 4 March 2024]; Instagram, 2023. DSA transparency report. [accessed 4 March 2024].

[250] TikTok, 2023. TikTok's DSA Transparency Report. [accessed 4 March 2024].

[251] Common Sense Media response to 2023 Ofcom Protection of Children Call for Evidence.

[252] Glitch response to 2023 Protection of Children Call for Evidence.

[253] Resolver, a Kroll business (formerly Crisp) meeting with Ofcom, 13 June 2023.

service, but that services should have regard to the particular needs of its UK user base as identified in its risk assessment in relation to language.

Meeting spikes in demand for content moderation driven by external events

16.182 We consider that for a content moderation function to be effective, services need to build in flexibility. In response to the 2022 Illegal Harms Call for Evidence, BSR (Business for Social Responsibility) stressed the importance of services "investing in the capability to scale-up/scale-down on short notice to respond to crisis events that can result in sudden spikes in illegal content."[254] We recommend that a similar approach should be taken regarding content that is harmful to children.

16.183 A study which analysed the content of messages shared on a forum found notable increases in the posting frequency on the forum following reports of celebrity deaths by suicide.[255] It also found that posts following celebrity deaths by suicide expressed greater negativity, raised cognitive bias, increased self-attentional focus and lowered social integration in the aftermath of celebrity deaths by suicides.[256] Evidence also suggests that suicide or self-harm content that is based on real events and challenges are likely to have a particularly detrimental impact on vulnerable users including children.[257] This suggests that it is important for services to consider how their content moderation systems deal with spikes in content that is harmful to children which is brought around by an external event, such as a celebrity death by suicide.

16.184 Following the start of the 2023 Israel-Gaza war, U2U services and other organisations reported an increase in harmful content online, including that which encourages hate and incites violence and graphic violent videos and images.[258] We are aware that some services have adjusted their content moderation capabilities in response to these events. For example, in response to the Israel-Gaza crisis, TikTok has added more moderators who speak Arabic and Hebrew to review content related to these events.[259] Meta has established a

---

[254] BSR response to 2022 Illegal Harms Call for Evidence.

[255] Kumar, M., Dredze, M., Coppersmith, G., & De Choudhury, M. 2015. Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides. HT ACM Conference on Hypertext and social media. September 2015. National Library of Medicine. [accessed 24 April 2024].

[256] Cognitive Attributes: Post-suicide content shows greater cognitive biases. Posts are less certain, show increased negation, and use more perception centric words, such as words in the category 'feel'. The psychology literature indicates such cognitive biases to be associated with lower emotional stability and increased self-consciousness.

[257] Though this challenge refers to a specific event, it shows that suicide or self-harm content based on real events is likely to have a particularly detrimental impact on vulnerable users. The study reviews the content of the posts about the challenge and suggests even where people are posting 'positive messages' (e.g. criticism of the challenge), the posts could contribute to social contagion and normalisation. Therefore, even where content is not directly promoting suicide or self-harm, there can still be a detrimental impact on vulnerable users. Khasawneh, A., et al. 2020. 'Examining the self-harm and suicide contagion effects of the Blue Whale Challenge on YouTube and Twitter: Qualitative Study', *JMIR Mental Health*, 7(6). [accessed 4 March 2024].

[258] Amnesty International, 2023. Global: Social media companies must step up crisis response on Israel-Palestine as online hate and censorship proliferate. 27 October 2023. [accessed 4 March 2024]; Scott, M., 'Graphic videos of Hamas attacks spread on X'. Politico, 9 October 2023. [accessed 4 March 2024]; Meta Oversight Board, 2023. 'Hostages Kidnapped from Israel'. [accessed 4 March 2024]; Meta Oversight Board, 2023. 'Al-Shifa Hospital'. [accessed 4 March 2024].

[259] TikTok, 2023. 'Our continued actions to protect the TikTok community during the Israel-Hamas war'. [accessed 4 March 2024].

special operations centre staffed with experts, including fluent Hebrew and Arabic speakers, with the aim of removing violating content faster.[260]

16.185 Information obtained from services' risk assessments, tracking evidence of new kinds of content that is harmful to children and other relevant sources of information,[261] could be used to understand where and when demands for harmful content might happen. In Volume 4, Section 11, we set out our reasons for proposing that all U2U services that are multi-risk for content harmful to children (regardless of size) and all large low-risk U2U services should track evidence of new kinds of content that is harmful to children on the service, and unusual increases in particular kinds of harmful content.

16.186 In instances where systems may need to deal with sudden harm events or spikes in harmful content, redeploying resource may draw resource away from another part of the system. Services that have plans in place to ensure that harmful content across the system is dealt with expeditiously are more likely to protect their users appropriately. Hence, we propose that services should consider the potential for spikes in problematic and potentially harmful content, or in other words the propensity for external events to lead to a significant increase in demand for content moderation on the service.

Rights assessment

16.187 This proposal recommends that services should ensure that content moderation functions are resourced so that performance targets are met (Measure CM3) and internal policies (Measure CM4) are followed. This proposed measure should therefore be seen as part of a package of measures relating to content moderation for content harmful to children, including Measures CM1 and CM2, for which we have assessed the rights impacts above.

Freedom of expression and association

16.188 We do not think that this proposal to ensure that content moderation functions are well resourced should have any specific adverse impacts on users' rights to freedom of expression or association. Instead, we consider that content moderation functions that are well resourced should result in more accurate decisions being made as staffing levels will mean that moderators are given the time to consider decisions properly and more swiftly, without the pressures and potential errors that under resourcing can bring. We therefore consider that recommending services have well-resourced content moderation functions is likely to assist in securing services moderate content swiftly and accurately, and therefore is more likely to safeguard against disproportionate impacts on users' rights to freedom of expression. If the result is that users, particularly children, are better protected from harm, it may also have a positive impact on children's freedom of expression and association as they may feel safer in using such services.

Privacy

16.189 We do not consider that our proposal that services ensure their content moderation functions are well resourced would have any specific adverse impacts on users' right to privacy beyond those already considered in connection with Measures CM1 and CM2 above. Instead, for the reasons set out above, we consider this proposal is likely to assist in securing

[260] Meta, 2023. 'Meta's Ongoing Efforts Regarding the Israel-Hamas War'. [accessed 4 March 2024].
[261] Under section 11(5) Online Safety Act, U2U services have a duty to notify Ofcom of non-designated content (NDC) identified in the Children's Risk Assessment.

services moderate content swiftly and accurately, and therefore in a way that safeguards against disproportionate impacts on users' rights to privacy.

Impacts on services

16.190 Service providers are expected to incur direct costs if they have to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

16.191 The total ongoing cost of resourcing services' content moderation functions in line with this measure is likely to be substantial, particularly for larger and riskier services with large volumes of relevant content to moderate. Whilst many services would in any case have some level of resource allocated to content moderation, a higher level of resources may be required to fully give effect to the policies and targets set out in Measures CM2, CM3 and CM4.

16.192 We expect that the level of resource required to implement the proposed measure will vary by size of service and also depend upon the policies they develop, and the nature and volume of harmful content present on their service. In general, we would expect costs to be higher for larger services, as larger services will tend to have a higher volume of content to review and therefore require more resource. For example, we understand that large services such as Meta and YouTube currently use upwards of 15,000 content moderators. At the same time, economies of scale are likely to mean that many smaller services face a higher moderation cost per user than large services. In any case, it is for services to consider the level and types of resource required to meet this measure, and to what extent this may entail additional resource and cost.

16.193 For providers of services likely to be accessed by children who are also in scope of the related measure proposed in our Illegal Harms Consultation, we consider that there may be some limited overlaps between the two measures.[262] For services which are already resourcing their content moderation systems in order to give effect to internal content policies and performance targets relating to illegal harms, these costs may be somewhat reduced in cases where there are synergies between the two types of content moderation, for example where a piece of content is both illegal content and content harmful to children. It is also possible that the same resources could be used to review both suspected illegal content and content harmful to children, which could help to manage costs in some cases (e.g. when there is a peak in prevalence of one particular kind of content).

16.194 In all cases, the magnitude of costs is likely to be further influenced by the type of detection and review processes used. Services will have flexibility over the mix of human and automated content moderation that they use.

16.195 For example, automating content moderation processes require both one-off infrastructure investment and different information and communication technology (ICT) professionals' time. Larger services may be able to develop these in house, but the costs of doing so can be high. Due to this, smaller services may outsource development to a third party or use off-

[262] See paragraph 12.147, Illegal Harms Consultation.

the-shelf third-party solutions.[263] In addition, system updates and licensing costs can be expensive and add to ongoing costs.

16.196 If content moderation primarily involves human moderators, resourcing costs will primarily depend on how many moderators are needed. In addition, for content moderation resources to be effective in meeting policies and targets, human moderators may require training (see Measure CM6 below).

Which providers we propose should implement this measure

16.197 This proposed measure is linked to and would be effective for those services which have content moderation policies and performance targets in accordance with Measures CM2, CM3 and CM4. This measure is important for those content moderation measures to have the intended effect.

16.198 Our analysis suggests that this measure could impose significant costs on services. However, we consider that where content moderation functions are well-resourced, this will deliver very significant and important benefits. We would expect this to result in a material reduction of harm to children compared to a counterfactual scenario where the service operates with a lower level of resources that may be insufficient to fully implement their internal moderation policies and achieve targets.

16.199 The costs of this measure are likely to scale with the benefits, as services with a higher risk of hosting content harmful to children will have a larger volume of content to review and therefore higher costs. However, the benefits of such content being identified and action taken regarding it will also be higher.

16.200 We propose to apply this measure to all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level). As we are proposing that Measures CM2, CM3 and CM4 would also apply to the same set of services, we consider it important that this measure apply to these services too, to ensure that those proposed measures are effective and able to reduce harm as discussed earlier in this section. We consider this proportionate for services that are multi-risk for content harmful to children given the risk of harm they pose to children, but also for large services that are not multi-risk for content harmful to children, as large services are typically more complex and may have a large volume of content moderation cases even if they are low-risk. We consider there is a material potential benefit from this measure, and its associated measure, even for such services, mitigating the risk of content moderation failures which could affect a large number of users, including children. Taking into account that large services will generally have greater capacity to resource their content moderation functions, we therefore consider it proportionate to apply this measure to all large services (as well as all multi-risk services).

16.201 This measure relates to resourcing well content moderation functions to give effect to measures CM2 to CM4, which do not extend at this stage to smaller services which are not multi-risk. For this reason, we also do not recommend this measure CM5 for smaller services which are not multi-risk. However, we note that these services should, in any case, ensure

[263] Pre-built solution offered by a third-party vendor.

that they have adequate resources to enable them to give effect under Measure CM1, even if we give more flexibility as to how they achieve that.

16.202 We are therefore proposing that this measure should apply to all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level).

Provisional conclusion

16.203 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm U2U services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the draft Children's Safety Codes. For the draft legal text for this measure, please see PCU B5 in Annex A7.

Measure CM6: Ensure content moderation teams are appropriately trained

Explanation of the measure

16.204 We propose that all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level) should provide training to content moderation teams (excluding volunteer moderators), having regard to factors including risk assessment information and evidence pertaining to emerging harms, as well as remedying gaps in content moderation staff's understanding of specific harms. We set out our proposed recommendation for providing materials to volunteer moderators in Measure CM7 below.

16.205 As set out in relation to Measure CM1 (have in place content moderation systems and processes), where a service has become aware of content that is harmful to children, they should swiftly action it in accordance with the children's safety duties and requirements in the Act. It follows that that the moderators carrying out this work need to know the relevant content policies that apply and how to carry out the relevant actions.

16.206 For services which are subject to Measure CM2 (set internal content moderation policies), Measure CM3 (setting targets) and Measure CM4 (adopting a prioritisation framework), we consider it very unlikely that it would be possible for moderators to give effect to such content moderation policies, targets and frameworks without training and additional materials such as: definitions and explanations around specific parts of the content moderation policy, enforcement guidelines, examples, and visuals of the tool or interface moderation staff will use to carry out their job.[264].

16.207 In this section, we are proposing recommending that services which have content moderation policies should ensure that people working on content moderation for children receive training (this may be in person, via online means or a hybrid approach) and materials that enable them to moderate content in accordance with their internal content moderation policies and therefore meet their set performance targets and other measures we propose in this section.

[264] Trust and Safety Professional Association, Setting Up a Content Moderator for Success. [accessed 11 March 2024].

Effectiveness at addressing risks to children

16.208 We know that many services already train their moderators and other relevant members of staff, or outsource to moderators and others who are trained, to identify and action content harmful to children, illegal or violative content, as well as providing supporting materials to help them do so.[265]

16.209 Several services told us they train their moderators to action content harmful to children and violative content and outlined (at a high-level) what kinds of training and support they receive. For example, some services (Meta, TikTok) told us that new hires in content moderation teams receive onboarding training before commencing their specific roles, which can include: training on specific policies, shadowing senior staff to understand how policies and procedures are applied in practice, and training on relevant systems.[266] These services also noted that they have on-going training, learning and development in place and that performance is assessed via exams.

16.210 Some services publicly outline what kinds of training and supporting materials they provide to their staff involved in content moderation. For example, Meta says its review teams "undergo extensive training to ensure that they have a strong grasp on our policies, the rationale behind our policies and how to apply our policies accurately".[267]

16.211 In response to the 2023 Protection of Children Call for Evidence, a number of civil society organisations, including 5Rights, Refuge, Glitch, Global Partners Digital, the South West Grid for Learning (SWGfL) and the Samaritans, stressed the importance of training.[268] The importance of training is also supported by broader academic and civil society literature and research.[269]

16.212 While services did not tell us exactly how often they train staff involved in moderation, several did say they trained their staff regularly (Roblox[270] and X (formally known as Twitter).[271]

16.213 In response to the 2022 Illegal Harms Call for Evidence, Global Partners Digital told us that services should provide regular training to moderators, "on the detail and application of the

---

[265] Morgan Lewis, 2023. Emerging Market Trend: An Overview of Content Moderation Outsourcing. [accessed 25 September 2023]; NYU Stern Center for Business and Human Rights, 2020. Who Moderates the Social Media Giants? A Call to End Outsourcing. [accessed 25 September 2023]

[266] Ofcom VSP information gathering from TikTok – 25/07/2022.

[267] Meta, 2022. How review teams are trained. [accessed 4 March 2023]

[268] 5Rights response to 2023 Protection of Children Call for Evidence; Refuge response to 2023 Protection of Children Call for Evidence; Glitch response to 2023 Protection of Children Call for Evidence; Global Partners Digital (GPD) response to 2023 Protection of Children Call for Evidence; SWGfL response to 2023 Protection of Children Call for Evidence; Samaritans response to 2023 Protection of Children Call for Evidence.

[269] Ofcom, 2019. USE OF AI IN ONLINE CONTENT MODERATION; The Alan Turing Institute, 2021. Understanding online hate: VSP Regulation and the broader context. [accessed 4 March 2023].

[270] Roblox response to 2022 Illegal Harms Call for Evidence.

[271] Twitter, 2023. "We have teams spread around the world specifically trained in this work so that we can provide this level of coverage in the languages we serve on Twitter" and "Updates about significant current events or rules and policy changes are shared with all content reviewers, to give guidance and facilitate balanced and informed decision making. In the case of rules and policy changes, all training materials and related documentation is updated." Twitter response to 2023 Ofcom Call for Evidence: Second phase of online safety regulation.

respective terms of service and ensuring that moderators are aware of any changes made ahead of their implementation".[272]

16.214 The Trust & Safety Professional Association states on its website that before launching a policy change, staff involved in content moderation need to be trained on the change. Services may choose to carry out the training in a number of ways, either by giving it directly themselves, through external trainers, and/or via e-learning. Lastly, the Trust & Safety Professional Association said that minor policy or processes changes may take place via communication, for self-learning, rather than through training refreshers.[273]

16.215 Some stakeholders responding to the 2023 Protection of Children Call for Evidence (Glitch and Global Partners Digital) also spoke about the importance of providing moderators with materials that support them in identifying and actioning content harmful to children.[274]

16.216 We understand that the people working in content moderation would mostly be staff employed or contracted by providers as dedicated content moderators. There may also be instances where it could include other roles in the business where specific expertise or advice is required such as: Trust and Safety staff; quality assurance and compliance staff; subject matter experts; lawyers and other legal staff; risk management staff; operations staff; engineers; and developers.[275]

16.217 We are aware that some services such as Discord, Freecycle, Nextdoor, Reddit, Twitch and WhatsApp use volunteers to help them moderate content (sometimes referred to as 'community-reliant' moderation).[276] While this measure does not include volunteer moderators due to a significant extra cost burden for services that we do not consider is justified by the benefit, we have considered an option that services that are multi-risk for content harmful to children should provide materials to volunteer moderators, see Measure CM7.

16.218 Specific materials provided to content moderators in scope of this proposed measure may include the content standards that fall under Measure CM1 of having content moderation and systems in place and Measure CM2 of setting internal policies but also include any other associated materials. They may also include definitions and explanations around specific parts of the policy, prioritisation frameworks, KPIs, enforcement guidelines, examples, and visuals of the review interface (i.e. the tool or interface moderation staff will use to carry out their job).[277] What is provided may vary depending on a number of factors, including, for example, the type of service, the type of content being moderated, and the local laws and regulations in the UK.

---

[272] Global Partner Digital (GPD) response to 2022 Illegal Harms Call for Evidence.
[273] Trust & Safety Professional Association. Setting Up a Content Moderator for Success. [accessed 4 March 2024].
[274] Global Partners Digital (GPD) response to 2023 Protection of Children Call for Evidence . Glitch response to 2023 Protection of Children Call for Evidence..
[275] Trust and Safety Professional Association. Key Functions and Roles. [accessed 24 March 2024]
[276] Discord, no date. Safety Library. [accessed 26 April 2023]; Freecycle, no date. Moderator Resources. [accessed 4 August 2023]; Freecycle, no date. New Moderator Orientation. [accessed 4 August 2023]; Nextdoor, no date. About Review Team members and moderation. [accessed 4 August 2023]; Reddit, no date. Reddit mods. [accessed 4 August 2023]; Twitch, no date. Guide for Moderators. [accessed 4 August 2023]; WhatsApp, no date. 101: Building a Safe Community. [accessed 4 August 2023].
[277] Hopkins, N., 2017. Revealed: Facebook's internal rulebook on sex, terrorism and violence, The Guardian, 21 May. [accessed 4 March 2024]; Trust & Safety Professional Association. Setting Up a Content Moderator for Success. [accessed 04 March 2024].

16.219 Based on the information above, we consider that training relevant staff involved in moderation, as well as providing them with relevant materials, is an important component of ensuring internal moderation processes are effective. We consider staff that have been trained on how to identify and action content harmful to children are more likely to be equipped with the knowledge and skills to do it when compared to those who are untrained.

16.220 We also think that relevant staff involved in moderation who are trained regularly will have up-to-date knowledge of content moderation policies, as well as on the systems they are using to carry out their job.

16.221 We do not intend to set any specific expectations around on how often training or supporting materials should be refreshed, as it may depend on a number of factors, including a person's role and performance. However, if moderators are trained on any major changes to policies or processes relating to content moderation and provided with new or updated supporting materials, they are more likely to be able to give effect to them accurately and consistently.

Factors to consider in the training of staff involved in content moderation and supporting materials

16.222 As set out above, we consider that service providers are best placed at present to determine what is appropriate for their services in terms of the detail of their training and materials. However, service providers that do not have regard to these possible factors are unlikely to have their content moderators trained appropriately to protect children. We therefore consider services should have regard to the below factors, when preparing and delivering content moderation training and materials.

Risk assessment and information pertaining to the tracking of signals of emerging harm:

16.223 A service's risk assessment will be one of the key sources of information telling a service provider what risk of content harmful to children they have on their service and will form the basis for internal content policies (see Measure CM2). As moderators should be focused on enforcing the internal content policies, training should also be focused on these policies.

16.224 In Governance and Accountability (Volume 4, Section 11), we are also consulting on a proposed recommendation that services should track signals of new and emerging harm – Measure GA5. If, following consultation, we remain of the view we should recommend this, this information would be one of the key sources of information about how content harmful to children manifests, and it is therefore crucial services use this to inform their content moderation training and supporting materials.

Remedying gaps in moderation staff's understanding of specific harms:

16.225 In response to the 2023 Protection of Children Call for Evidence, a few services discussed specialist training, including for specific harms. For example, Patreon said that it provided vertical-specific moderator training for each of their policy vertical areas e.g. minor safety, sexually graphic content, and hate speech.[278] Google said that "moderators receive regular training, including to identify content that is harmful to children, in line with our policies on harmful or dangerous content, harassment and cyberbullying".[279] We also know that many services, particularly larger ones, give their staff involved in moderation specialist training

[278] Patreon response to 2023 Protection of Children Call for Evidence.
[279] Google response to 2023 Protection of Children Call for Evidence..

and materials in particular areas, including illegal harms, other harms, freedom of expression, and user rights.[280]

16.226 Several civil society organisations recommended specialist training on specific harm areas (Global Partners Digital[281]), including, gender-based violence (Glitch[282]); child safeguarding, risks to children, and knowledge of child development (5Rights[283]); and awareness of learning disabilities (MENCAP[284]). Global Partners Digital also stressed the importance of training moderators in the potential impact to users' rights and freedom of expression.[285]

16.227 There may be occasions where harms-specific training and materials can be helpful in identifying and removing harmful content due to the unique, complex, novel or serious nature of a given harm, or because certain harm or harms, for example self-harm, may be particularly prevalent on a service and so require more in-depth understanding.[286] For example, although some suicide, self-harm and eating disorder content can be easily identified as harmful content, it can be difficult for content moderators to determine whether a user's personal account of an eating disorder depicts a person speaking about their recovery or if it encourages, promotes or provides instruction. If training materials are given to moderators where a service has identified a gap in moderators' understanding of a specific harm, and where they deem there to be a specific risk, this should improve the effectiveness of content moderation and therefore children should encounter less harmful content. Services may also refer to Ofcom's Guidance on content harmful to children (Volume 3, Section 8) as a resource in remedying gaps in moderator training.

16.228 We do not consider that it would be appropriate to specify in Codes how often materials should be revised, or training should be redelivered. However, services should take their ongoing children's access assessment and children's risk assessments duties, including changes on the service, into consideration, when reviewing their policies and their training to ensure that risk is captured. A service which failed to refresh training and materials

---

[280] Ofcom VSP information gathering from TikTok – 25/07/2022.

[281] Global Partners Digital said that "content moderators should also be able to specialise and progress in expertise on a particular content type". Global Partners Digital response to 2023 Protection of Children Call for Evidence.

[282] Glitch said there should be "comprehensive" training for moderators on "online gender-based violence and different tactics of online abuse, and how abuse specifically targets women, Black and minoritized communities and users with intersecting identities". Glitch response to 2023 Protection of Children Call for Evidence.

[283] 5Rights commented that human moderators should receive training in how to identify risks to child safety, "including knowledge of risks to different groups of children and the full range of content and activity that is illegal or might be harmful to a child. This also includes knowledge of the stages of child development and awareness of how children's capacities, vulnerabilities and behaviour change as they grow. 5Rights response to 2023 Protection of Children Call for Evidence.

[284] MENCAP said that to moderate content more accurately, there should be "awareness training to moderators on learning disability as well as other groups deemed more likely to be subjected to online harms and illegal content". MENCAP response to 2022 Illegal Harms Call for Evidence.

[285] Global Partners Digital response to 2022 Illegal Harms Call for Evidence. Similar comments regarding moderation and user rights and freedom of expression were raised in its response to Ofcom's 2023 Protection of Children Call for Evidence, particular concerns were raised about automation bias.

[286] This article identified that one of the barriers to moderating self-harm content was "vagueness within the guidelines". Moderators of online forums said moderation was made easier when all staff had 'the same understanding of the guidelines' to keep forums "safe and providing consistency". Perowne, R. and Gutman, L.M. (2022) 'Barriers and enablers to the moderation of self-harm content for a young person's online forum', *Journal of Mental Health*, pp. 1–9.

following any major changes to policies or processes relating to content moderation that is to do with content that is likely to be harmful to children would not be enabling its moderators to moderate content in accordance with Measures CM1-5 above.

Other issues to note

16.229 A number of civil society respondents to the 2023 Protection of Children Call for Evidence stressed the importance of supporting the wellbeing of staff involved in content moderation, including the Samaritans, 5Rights, Glitch and Molly Rose Foundation.[287] Global Partners Digital noted that adequate financial, emotional and psychological support is "vital to reduce turnover and burnout in content moderation teams, which limits institutional knowledge and consistency between decisions and lowers the overall accuracy of the content moderation systems".[288]

16.230 Research suggests that human content moderation has the potential to cause significant impacts on the wellbeing of staff members, including secondary trauma, altered psychological wellness, content fatigue and burnout.[289] Some providers offer controls to moderators when reviewing content, such as applying blurring or audio removal, though this is not universal.[290] Some providers also have wellbeing support in place for moderators such as counselling and mental health support, such as Twitter[291], Patreon[292] and Google. [293]

16.231 We recognise the significant impact that human moderation of content can have on the wellbeing of an individual and the importance of providing appropriate supervision and support in this area. However, we acknowledge that the responsibility towards employed moderators is within the employers' remit and therefore would only be relevant to our remit if it impacted on user safety. We welcome evidence from stakeholders on this, to which we would have regard in planning our work on future iterations of our Codes.

Rights assessment

16.232 This proposal recommends that services provide appropriate training to content moderation staff to help ensure that their internal policies (CM2) are followed, performance targets are met (Measure CM3) and prioritisation framework (Measure CM4) is followed. This proposed measure should therefore be seen as part of a package of measures relating to content

---

[287] Samaritans' response to 2023 Protection of Children Call for Evidence; 5Rights response to 2023 Protection of Children Call for Evidence; Glitch response to 2023 Protection of Children Call for Evidence; Molly Rose Foundation response to 2023 Protection of Children Call for Evidence.

[288] Global Partner Digital response to 2023 Protection of Children Call for Evidence.

[289] Steiger, M., Bharucha, J.T., Venkatagiri, S., Martin J. Riedl, J.M., and Lease, M., 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. [accessed 4 March 2024].

[290] The British Psychological Society, 2022. Invisible workers, hidden dangers. 22 April [accessed 14 September 2023].

[291] "We have a full suite of support services available for our employees, including content moderators. Some of the measures we take include, establishing resiliency programs across all of our partners, committing to recurring leadership visits and ongoing feedback loops and communications between all of our teams". Twitter response to 2023 Protection of Children Call for Evidence.

[292] "Moderators are also afforded access to individual and group wellness sessions to help build resilience and assist with processing difficult and disturbing content." Patreon response to 2023 Protection of Children Call for Evidence.

[293] Google response to 2023 Protection of Children Call for Evidence.

moderation for content harmful to children, including Measures CM1, CM2, CM3 and CM4, for which we have assessed the rights impacts above.

Freedom of expression and association

16.233 We would not expect this proposal to have any specific adverse impacts on users' rights to freedom of expression or association. Instead, we consider that providing appropriate training to content moderation teams should result in more accurate decisions being made, as staff should be more aware of which content should be actioned and how to do so. This is therefore likely to assist in ensuring services moderate content in a way that safeguards against disproportionate impacts on users' (including both children's and adults') rights to freedom of expression. If the result is that users, particularly children, are better protected from harm, it may also have a positive impact on children's freedom of expression and association as they may feel safer using such services. Adult users will also benefit from this proposed measure as content would be accessed and shared appropriately if fewer errors are made and decisions around content moderation are actioned more swiftly.

Privacy

16.234 We would not expect this proposal to have any specific adverse impacts on users' rights to privacy. We consider that the training of content moderators would help safeguard users' privacy as staff would be clear on which content should be actioned, thereby resulting in more accurate decisions being made. This should reduce the likelihood of inaccurate personal data and ensure a degree of fairness in the processing of that personal data. It will also mean that content moderators understand which content should be actioned and how, thereby enabling consistency in decision making. We consider this is likely to result in a more proportionate approach to content moderation and therefore likely to safeguard users' rights.

Impacts on services

16.235 Service providers are expected to incur direct costs if they need to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

16.236 In order for a service provider to implement the measure, it would incur two main types of cost. Firstly, the costs of developing the training material, both upfront costs and ongoing costs of keeping this updated. The second is the cost of delivering the training to moderators. Services, which are not within scope of the related measure proposed in our Illegal Harms Consultation and do not otherwise already have parts of this measure in place, would incur the full costs of developing the training material, which we discuss below.

16.237 The costs associated with delivering the training to content moderators will be impacted by the format of training chosen (e.g. delivered by a human trainer each time or via a video/interactive interface, or on-the-job training) and will also depend upon the number of staff to be trained and the duration of the training. We assume that content moderators will not be available to perform their usual role during the training process, but will be compensated during the training process.

16.238 The duration of the training needed will usually to be longer the more complex and diverse the range of possible harmful content is on a service. As an indicative estimate, we assume a

range of two to six weeks duration for someone having this training for the first time.[294] Based on this duration and a range for pay, we estimate that the costs of providing training for one new content moderator could be between £3,000 and £18,000, and for a new software engineer[295] between £5,000 and £28,000.[296] This includes both the wage cost of the employees being trained plus an uplift to capture the costs of preparing and delivering the training. If content moderators are based in countries with lower labour costs than the UK, then the lower end of the wage range we have assumed will overstate the costs. These costs may also vary depending on whether the training is by in-house staff or by an external provider.

16.239 In addition to these costs of training new content moderators and software engineers, there will be some ongoing costs for refresher training and training in new harms on the services. We expect the annual costs of these to be lower.

16.240 For providers of services likely to be accessed by children who are also in scope of the related measure proposed in our Illegal Harms Consultation, we consider that there may be some limited overlaps between the two measures.

16.241 In terms of developing the training material, whilst the types of harms and associated content are not the same, and services may need to make changes to training content and duration in order to comply with the CSD so that training is adequate for the OS regime, there is likely to be a limited degree of overlap in the training content required for the two types of material. We therefore expect that providers who already have training in place to cover these harms will have slightly lower costs as a result of this measure than those who have no training in place for content moderators at all.

16.242 All other things being equal, smaller services will have fewer content to review, smaller content moderator teams and therefore will incur lower costs of training. While costs for services will scale with the risk of harm, this will come with a proportionate benefit. In general terms, we would expect costs to vary with the potential benefits, in that services with higher risk of hosting content harmful to children are likely to need more content moderators and require them to be trained on more harms, therefore resulting in higher training costs. Conversely however, these services are likely to have more content harmful to children and therefore higher benefits from having well trained content moderators who can take action effectively regarding these kinds of content.

16.243 These costs are also mitigated by the fact that this measure does not specify exactly how services should provide training to content moderators, giving services some flexibility in what they do. Services can decide the most appropriate and proportionate approach to

---

[294] This range is consistent with examples we are aware of from the industry, suggesting that in most cases the relevant training period could be shorter than six weeks. Equally, where content moderation systems are particularly complex or a service faces a multitude of risks across different kinds of CHC and different types of media, it remains possible that more extensive training could be appropriate for some staff.

[295] We would expect that only ICT colleagues directly involved in operationalising content moderation systems would need to do this training.

[296] This is based on our assumptions on wage rates set out in Annex 12. We also assume that the wage cost of the people being trained represents only half of the total costs of the training. This is consistent with the Department for Education estimation that the wage cost of staff being trained accounted for about half of all training expenditure in 2019, although this varies by the size of the firm and the sector. We assume this excludes the 22% uplift that we have assumed elsewhere for non-wage labour costs, so we have not also increased these wages by 22%. Source: Department for Education (DfE), 2020. Employer Skills Survey 2019: Training and Workforce Development, pp38-40. [accessed 5 February 2024].

training content moderators for their own contexts. This flexibility provides a cost-effective and proportionate approach for each service.

Which providers we propose should implement this measure

16.244 This proposed measure is linked to and would be effective for those services which should have content moderation policies in accordance with Measure CM2, set performance targets in accordance with Measure CM3 and have policies on prioritisation in accordance with Measure CM4. We consider that moderators need to be appropriately trained in order to give effect to those measures. We recommend this proposed Measure CM6 for the same services in scope of Measures CM2-4.

16.245 We consider the benefits of this measure are likely to be high. This is because content moderator training is important in implementing effectively a service's content moderation policies to reduce harm and comply with its duties. Well-trained and prepared paid content moderators are more likely to be able to identify content harmful to children and, under the service's content standards apply the correct action to take, thereby reducing the harms that may result from such content. As the number of content moderators that need training is likely to depend upon the size of the service and the volume of content that needs to be assessed, the costs of this measure are likely to scale with the benefits. As such, this measure is likely to be proportionate for services that are multi-risk for content harmful to children given the potential for harm to children on such services.

16.246 As per Measures CM2, CM3 and CM4, we also consider the Measure CM5 is proportionate for large services that are not multi-risk for content harmful to children. Large services are typically more complex and may have a large volume of content moderation cases even if there is low-risk. We consider there is a material potential benefit from appropriate training under this measure, even for such services, mitigating the risk of content moderation failures which could affect a large number of users, including children. The training may also promote consistency in approach where a service has many moderators, which may be the case on a large service even if low-risk. We also note that large services are likely to have sufficient resources to train moderators in line with the proposed measure.

16.247 At this stage we do not consider it proportionate to recommend this measure to smaller services that are not multi-risk for content harmful to children, as these services are likely to moderate lower volumes of content that may be harmful to children and the benefits are therefore likely to be lower. It is likely that these services will still need to consider appropriate steps to equip content moderation staff to be able to implement Measure CM1. However, for such services we are not recommending formal training with the specific elements set out in this measure, thereby providing more flexibility to such services.

16.248 Many services also use volunteers to help them moderate content in addition to paid moderators. Volunteer moderators are not in scope of this measure, but are addressed separately in Measure CM7.

16.249 We are therefore proposing that this measure should apply to all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level).

Provisional conclusion

16.250 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm U2U services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the draft Children's Safety Codes. For the draft legal text for this measure, please see PCU B6 in Annex A7.

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

Explanation of the measure

16.251 We are aware that many services currently use volunteer moderators (sometimes referred to as community moderators) to moderate content and that on these services, volunteer moderators often perform the significant proportion of moderation action.[297] We do not anticipate that the services this proposed measure applies to will rely on volunteer moderation alone due to the size of the service and the degree of risk that content harmful to children is present on that service. We propose that all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level) should provide appropriate materials to volunteer moderators for their roles.

16.252 Volunteer moderators that have access to appropriate materials are more likely to carry out their roles effectively.

16.253 We considered the inclusion of volunteer moderators in Measure CM6 which sets out our proposal that all U2U services likely to be accessed by children that are multi-risk for content harmful to children (regardless of size) and all large U2U services (regardless of risk level) should ensure their paid content moderation teams are appropriately trained. However, we considered there would be a significant extra cost burden for services if we were to propose that all volunteer moderators are trained under that measure. At this time, Measure CM6 is limited to paid moderators with Measure CM7 working to ensure that volunteer moderators still have access to appropriate training materials.

16.254 As with other measures, we do not propose to be prescriptive about the form these materials should take, leaving scope and flexibility for services to tailor these resources according to their individual needs, so long as the contents of the resources enable volunteer moderators to fulfil their role in moderating content in accordance with Measure CM1 and CM2.

[297] For example, Reddit's 2022 Transparency report shows that 58% of content removed from Reddit was actioned by volunteer moderators. The total volume of removals by moderators in 2022 increased by 4.7% compared to 2021. We note that not all content actioned by volunteer moderators may be harmful, content may be violative of community rules. For example, not concerning the topic of the community. Reddit, 2022. Transparency Report. [accessed 4 March 2024]; Nextdoor's 2022 Transparency Report shows that volunteer moderators reviewed 92% of all reported content. Nextdoor 2022 Transparency Report.

16.255 Some services already provide some materials to volunteer moderators. For example, Reddit provides a moderator help centre for its moderators, including courses for moderators.[298] Discord offers tools, resources and guidance as part of its 'Safety Library'.[299] Twitch provides various information pages on topics such as "Guide for Moderators", "Combating Targeted Attacks" and "Managing Harassment" to aid moderators. Wikipedia provides pages on standards requirements. [300]

16.256 Evidence shows that children are present on services such as Twitch, Discord, Reddit and Snapchat,[301] all of which employ some form of volunteer moderation. In its response to the 2023 Call for Evidence, The Internet Commission also noted the presence of children on services that use both volunteer moderation and paid moderation, the need to provide training and support to moderators and how 'the layered approach of internal and community enforcement must operate coherently'.[302]

16.257 Further, our evidence shows harmful content manifests in different communities, groups, discussion forums, chat rooms etc. that children are likely to access. Evidence shows that groups and communities are a pathway to content harmful to children.[303] A study by Internet Matters noted that the users of chatrooms and forums were significantly more likely to experience all of the categorised online harms from their study, compared to users of other online activities.[304] We consider that providing materials to volunteer moderators could mitigate the risk that children encounter harmful content in these groups and communities. At the same time, it would allow children to continue to participate in communities that might appeal to them and from which they might benefit, such as sports, animals, fandoms and celebrities.

---

[298] Reddit provide a moderator help center. This outlines the basics of starting a community on Reddit, overview and explanation of individual moderation tools, community engagement and advice and materials. It also provides subreddits for news, support and requests. Reddit also provide volunteer moderators 'Reddit Mod Education Courses'. The layout and set up uses the way that the platform operates to provide materials and support to its users.

[299] Discord provides users further information and links to support on their 'safety and moderation' page, including information on how to develop server rules, links to moderation and community support to manage servers and "handling difficult scenarios as an Admin".

[300] Twitch provide various pages using pictures and videos to show the moderator's view of the channel and their tools.; Wikipedia's moderator access is dependent on hierarchy and are required to follow extensive procedural rules. Wikimedia, 2020. How Content Moderation and Anti-Vandalism Works on Wikipedia. [accessed 4 March 2024]; Oz, A., 2009. '"Move along now, nothing to see here": The private discussion spheres of Wikipedia', SSRN. [accessed 4 March 2024].

[301] Different research sources tracking children's use of these platforms include Ofcom's VSP, Online Experiences and Media Literacy trackers, with the latter showing that: 8% of 3-17 year old children use Twitch, 4% use Reddit, 9% use Discord and 46% of 3-17 year old children use Snapchat; see Ofcom, 2023 Children and Parents: Media Use and Attitudes (data tables here, table 31). Note: Snapchat allows moderators to be appointed for shared stories.

[302] The Internet Commission response to the 2023 Protection of Children Call for Evidence.

[303] Graphika presentation to Ofcom, 7 July 2023.

[304] The other harms listed were: "come across violent content; online bullying from people known; online bullying from strangers; come across sexual content; come across promoting dangerous eating habits; come across self-harm". The other online activities listed were: "broadcast videos streamed live; watch videos streamed live; play games against each other (multi-player); play or use software in the metaverse; upload or share videos they've made themselves; use messaging apps; use social media services". Internet Matters, 2023. Exploring the impacts of online harms. [accessed 28 March 2024].

16.258 Further, we are aware that children who self-harm[305], have an eating disorder or suicide ideation[306] often seek support in communities and groups online.[307] We consider this measure could mitigate the risk of children encountering PPC when looking for supportive content by providing volunteer moderators with materials to help them identify content, information about how to action the content appropriately and when to escalate content or situations to the service. Materials should be provided for volunteer moderators, to ensure a greater understanding of content harmful to children by volunteers and how that content should be actioned as per the service's terms of service. This would help services prevent children from encountering PPC and protect children from encountering PC and NDC, providing a safer online experience for children who engage in communities and groups.

16.259 Therefore, we are proposing recommending that services which use volunteer moderation as a form of content moderation should ensure that volunteer moderators are provided with materials that enable them to fulfil their role. This should support volunteer moderators to be able to identify and action content harmful to children in line with Measure CM1 (have in place content moderation systems and processes) and CM2 (internal policies) and, in turn, minimise the risk to children of encountering harmful content in such spaces.

16.260 Below we set out examples of circumstances in which volunteer moderators may benefit from receiving materials from the service provider. These include but are not limited to:

- where an increased risk to children encountering harmful content is identified by volunteer moderators who notice an increase in content harmful to children or an attempt by bad actors to take over a community by posting content harmful to children in their communities;
- volunteer moderators have escalated content as they are unsure what action to take to address certain content harmful to children or if they are uncertain whether content is PPC, PC or NDC;
- the creator of a community needs support with matters relating to appointing volunteer moderators including the number of volunteer moderators that is appropriate for a community of their size; [308]
- and/or[309] technical support is needed, where the volunteer moderation tools are not working.

16.261 We understand "community" and "volunteer moderation" to mean the following:

Definition box 1: What is volunteer moderation?

[305] Children and young people also joined communities and made online friends, who shared pro-self-harm content amongst themselves. Ecorys, 2022. Qualitative research project to investigate the impact of online harms on children. [accessed 4 March 2024].

[306] Another potential online source of suicide stories is the widespread availability of sites devoted to discussions about specific topics. These sites include discussion forums and boards as well as self-help venues where users can post questions and obtain help and reactions from others with similar interests. For example, Reddit, has a specific section dedicated to discussions about suicide. [accessed 25 March 2024]; Dunlop, S.M., More, E. and Romer, D. (2011) 'Where do youth learn about suicides on the internet, and what influence does this have on suicidal ideation?'. Journal of Child Psychology and Psychiatry, 52(10). [accessed 04 March 2024].

[307] "If you do self-harm or have an eating disorder, I think, you know, those communities are easy to find online, some supportive, and some more enabling". SEND Professional. Ecorys, 2022. Qualitative research project to investigate the impact of online harms on children. [accessed 04 March 2024].

[308] Twitch, Building a Moderation Team. [accessed 4 March 2024].

[309] Twitch, Building a Moderation Team (twitch.tv). [accessed 04 March 2024].

Community: "Community", also referred to as "groups" or "forum groups" refer to a user-to-user service functionality allowing users to create online spaces that are often devoted to sharing content on a particular topic. User groups can be open to the public or closed to the public, requiring a registered account and an invitation or approval from existing members to gain access.

"Volunteer Moderation", also referred to as "Community-reliant Moderation" and "Distributed moderation" typically refers to a form of moderation that combines formal policy made at the service level with community-specific rules by volunteer moderators at community level. This form of moderation relies on community members moderating content that does not align with community expectations. Volunteer moderation is often used as one type of moderation within a wider system.[310] For example, service providers may use volunteer moderators while also using pre- and post-moderation systems for certain types of content, or auto moderation features which allow users to set up rules, for example providing a definition and outline of rule structure in relation to automated moderation of profanity and slurs.[311]

Though many online services rely on users to aid in the process of moderation – primarily by flagging content for review – some providers rely on volunteer moderators much more substantially. These services may separate powers between the parent organisation and its subcommunities, with the parent organisation setting overarching norms and standards, which can be added onto by subcommunities contained within the site. This can be compared to a "federal system", as described by a Reddit representative, with baseline site-wide rules that must be obeyed by smaller subcommunities but can also be extended according to the discretion of sub-community moderator.[312]

Materials for volunteer moderators

16.262 Services should provide materials for their volunteer moderators that are relevant to the service and updated as necessary. The materials may be provided in different forms, depending on the service and could include, among other things, online modules, written resources and videos. The materials should provide appropriate information for volunteer moderators to be able to carry out their roles in relation to the protection of children from harmful content.

16.263 These materials should provide owners and moderators of communities and groups with an overview of their role and responsibilities as content moderators. Services may consider creating a Code of Conduct so that volunteer moderators understand what is expected of them.[313]

16.264 These materials should provide the necessary information to help volunteer moderators carry out their roles in relation to content harmful to children.[314] This may include, but is not limited to, the service's community guidelines and other content policies. It may also include

[310] Caplan, R., 2018. Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches. [accessed 4 March 2024].
[311] Discord, Developer Portal. [accessed 10 December 2023].
[312] Reference to "federal system" by a Reddit representative refers to a comparison to the US system of national and state governance. Caplan, R., 2018.
[313] Reddit provide volunteer moderators with a Moderator Code of Conduct. [accessed 25 April 2024]; Nextdoor provide a Nextdoor Community Programs Code of Conduct. [accessed 25 April 2025].
[314] See, for example, Perowne, R. and Gutman, L.M.,2022. Barriers and enablers to the moderation of self-harm content for a young person's online forum. *Journal of Mental Health*. [accessed 06 March 2024].

Ofcom's Guidance on Content Harmful to Children, unless already reflected in the services' own content policies.

16.265 In order to enable volunteer moderators to carry out their roles, materials would also need to provide information about the tools available to them to help them moderate content harmful to children, what actions they should take when moderating this content, and what the service's appeal process is.[315] Moderators should also be made aware of how and when they should escalate content for further moderation, as well as the wider content moderation systems and process within which they operate.

16.266 Services should also provide guidance, for example, on who can be designated as volunteer moderators and the importance of having enough moderators for the size and engagement of the community.

16.267 As with Measure CM6, which sets out that paid content moderation teams should be appropriately trained, there is no set best practice on how often materials should be refreshed or updated, however, where there are any major changes to policies or processes relating to content moderation relevant to volunteer moderators, volunteer moderators should be provided with new or updated materials.

16.268 Service providers should also ensure that the materials are clearly labelled and easily accessible, so that volunteer moderators are aware of their availability. We consider there are a number of ways services may wish to do this, for example:

- Sending communications to volunteer moderators about the materials when they become volunteer moderators.
- Where services have a code of conduct or dedicated area of their service for volunteer moderators, ensuring that the materials are easily discoverable and available.
- Sending reminders and nudges to volunteer moderators to remind them of the materials available.
- Where services have regular communications sent to volunteer moderators e.g. weekly newsletters, ensuring that how they can access the materials is clearly signposted.
- When volunteer moderators contact the service, reminding them of the materials available.

Rights assessment

16.269 This proposal recommends services in scope make materials available for volunteer moderators.

Freedom of expression and association

16.270 We would not expect this proposal to have specific adverse impacts on users' rights to freedom of expression and association. Volunteer moderators that are able to access material to enable them to carry out their roles should be likely to make more accurate decisions as they should be more aware of which content should be actioned and how. However, we acknowledge that services may potentially have less oversight and influence

[315] Cullen, A.L., and Kairam, S.R., 2022. Practicing moderation: Community moderation as reflective practice, Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1). [accessed 4 March 2024].

over volunteer moderators who are not employees or contracted staff because they may not have written contracts in place. We do not anticipate that the services this proposed measure applies to will rely on volunteer moderation alone due to the size of the service and degree of risk of content harmful to children being present on that service. It will also mean that volunteer moderators understand which content should be actioned, enabling some consistency in decision making. This will mean that decisions are likely to result in a more proportionate approach to content moderation by the service, and therefore likely to safeguard users' rights.

Privacy

16.271 We would not expect this proposal to have specific adverse impacts on users' rights to privacy. We consider that the provision of materials to moderators would help to safeguard users' privacy as they would have more information on which content should be actioned with support available from the service for any queries, resulting in more accurate decisions being made. However, we acknowledge that services may potentially have less oversight and influence over volunteer moderators who are not employees or contracted staff because they may not have written contracts in place. We do not anticipate that the services this proposed measure applies to will rely on volunteer moderation alone due to the size of the service and the degree of risk that content harmful to children is present on that service. All services should ensure they comply with data protection laws and consider relevant guidance from the ICO.[316] Such compliance should reduce the likelihood of inaccurate personal data and ensure fairness in the processing of that personal data. It will also mean that volunteer moderators understand which content should be actioned, enabling some consistency in decision making. We consider this is likely to result in a more proportionate approach to content moderation and therefore tend to safeguard users' rights.

Impacts on services

16.272 Service providers are expected to incur direct costs if they need to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

16.273 For a service provider to implement this measure, there would be an initial cost of creating or providing the materials. This could be done either internally if they have the relevant expertise, or externally. Where a service chooses to source these materials externally, the cost of this would depend on whether the service already employs external organisations to provide materials for human moderators.

16.274 We expect that the majority of services within scope of this measure would also be within scope of Measure CM6 (provision of training for paid moderators) and could therefore build on or adapt the materials developed for this measure. There may be small additional costs associated with this e.g. adapting the format of the materials so that they can be accessed online rather than in person, making the materials searchable, or adjusting the level of detail so that the materials are relevant for the role of a volunteer moderator on that particular service. However, we do not anticipate that these costs are likely to exceed a few thousand pounds.

[316] See from the ICO: Children's Code guidance and resources; A guide to the data protection principles; and Content moderation and data protection. [accessed 24 April 2024].

16.275 For services that rely only on volunteer moderators and do not have relevant existing materials developed for paid moderators, costs will be higher. However, we consider that the costs for these services associated with Measure CM7 would be considerably less than the cost estimates relating to providing training for paid content moderators that we outline in relation to Measure CM6 as these figures also include wage costs for moderators while receiving the training, which will not be incurred for volunteer moderators.

16.276 There would also be an ongoing cost to all services of updating the materials to ensure that they remain relevant. Even if it were possible to source an 'off the shelf' version of the materials, this would need to be updated and regularly reviewed in light of new and emerging harms.

16.277 The costs above are mitigated by the fact that this measure does not specify exactly how services should provide materials to volunteer moderators, giving services some flexibility to decide the most appropriate and proportionate approach for their own contexts.

Which providers we propose should implement this measure

16.278 As set out above, we consider that this measure could substantially reduce the risk of children encountering harmful content in online communities such as groups, discussion forums and chatrooms by ensuring that volunteer moderators are better equipped to deal with harmful content where it is detected in these communities.

16.279 We expect that the costs for large services or services that are multi-risk for content harmful to children of implementing this measure will generally be small, given that we expect the majority of these services to also be within scope of measure CM6 (paid content moderation teams are appropriately trained) meaning that they will already have in place materials for paid moderators in most cases. The additional cost of making some of that material available to volunteer moderators should not be too substantial. The costs of this measure are also likely to be lower for smaller, less complex services with fewer risks, for example because their volunteer moderators are dealing with less complex issues and fewer kinds of harmful content.

16.280 For any services that rely only on volunteer moderators and do not have existing resources developed for paid moderators that can be adapted, costs will be higher. However, we consider the benefits from implementing the measure will also be higher, as having well-informed and well-prepared volunteer moderators will be particularly important where services place greater reliance on these moderators, to reduce the risk of moderation failures that expose children to harm.

16.281 For any smaller services that are not multi-risk for content harmful to children and use volunteer moderators, they would still be expected to consider how to provide relevant information to volunteer moderators if appropriate as part of implementing Measure CM1, but we are not necessarily recommending they follow the specific approach to providing materials as described in this Measure CM7. Benefits from the measure would be lower on these services, given the more limited risk to children. On the other hand, the costs of this measure would be substantial for smaller services that are not multi-risk for content harmful to children, as they would not be in scope of Measure CM6 and therefore may not have existing materials that can be adapted. Therefore, at this time we do not consider that it would be proportionate to recommend this measure for such services.

16.282 We are therefore proposing that this measure should apply to all U2U services likely to be accessed by children that use volunteer moderation and are multi-risk for content harmful to children (regardless of size) and all large U2U services that use volunteer moderation (regardless of risk level).

Other options considered

Whether training for volunteer moderators should be mandatory

16.283 We considered the option of recommending that services ensure that training is completed in full by current and future moderators of groups and communities before individuals are permitted to moderate content.

16.284 Though services should consider ways to encourage volunteer moderators to use materials provided, we propose that training should not be mandatory. We recognise that recommending mandatory training may be disproportionate considering the number of volunteer moderators and the ease with which users can create communities and become volunteer moderators.[317] We also considered the additional burden on services of logging and recording training completed by volunteers, and the practical difficulties of ensuring that volunteers complete certain actions, which may not be feasible given that these individuals are not contracted to the provider.

16.285 Further, we considered the impact this added condition could have, firstly, on the ability of users to create communities and engage in real time with issues, for example, communities being set up in response to world events. Secondly, we considered its impact on participation, as the condition could discourage individuals from taking on the role of volunteer moderators and ultimately make content moderation less effective, leading to worse outcomes for users of the service.

16.286 However, we recognise the limited impact this measure may have if volunteer moderators do not utilise the training resources. Therefore, we welcome evidence on this for consideration in planning our work on future iterations of our Code.

Provisional conclusion

16.287 Given the harms this measure seeks to mitigate in respect of PPC, PC, and NDC as well as the risks of cumulative harm U2U services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU B7 in Annex A7.

[317] Li, H., Hecht, B., & Chancellor, S., 2022. Measuring the Monetary Value of Online Volunteer Work, Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 596-606. based on a 2022 study of 21,522 active Reddit volunteer moderators. [accessed 04 March 224]; Nextdoor enhanced its unique volunteer moderation model to support 210,900 volunteer moderators. Nextdoor 2022 Transparency Report. [accessed 04 March 2024].

New Measure (Illegal Content Code): If volunteer moderation is used, provide moderators with materials for their roles

Explanation of the measure

16.288 As set out above, we are aware that many services currently use volunteer moderators (sometimes referred to as community moderators) to moderate content. This would also include illegal content. We consider that if volunteer moderators have access to appropriate materials, they are more likely to carry out their roles effectively. However, we do not anticipate that the services this proposed measure applies to will rely on volunteer moderation alone due to the size of the service and the degree of risk of different kinds of illegal content. We propose that all U2U services that use volunteer moderation and are multi-risk (regardless of size) and all large U2U services that use volunteer moderation (regardless of risk level))[318] should provide appropriate materials to volunteer moderators for their roles.

16.289 As with other measures we have proposed in the draft Illegal Harms Consultation, we do not propose to be prescriptive about the form these materials should take. We leave scope and flexibility for services to tailor these resources according to their individual needs, so long as the contents of the resources enable volunteer moderators to fulfil their role in ensuring the provider to moderate content in accordance with Measures 4A and 4B in our draft Illegal Content Codes.

16.290 We are aware that many services currently use volunteer moderators to moderate content and that, on these services, volunteer moderators often perform the significant proportion of moderation action.[319] In response to the 2022 Illegal harms Call for Evidence, Mumsnet reported that it has a team of 14 freelance moderators and two staff members, who are on duty seven days a week. Additionally, Nextdoor has volunteer moderators on Neighbourhood Teams who are monitoring community discussions 24/7. Wikimedia also uses volunteer moderation and stated in its call for evidence response that 'content moderation on Wikipedia, and other volunteer-run free knowledge projects that the Foundation hosts and supports, is largely conducted by a community of nearly 300,000 global volunteer contributors'.[320]

[318] As per our proposed definition of 'multi-risk' in the Illegal Harms Consultation, where multi-risk means high or medium risk for at least two kinds of illegal harm.

[319] For example, Reddit's 2022 Transparency report shows that 58% of content removed from Reddit was actioned by community moderators. The total volume of removals by moderators in 2022 increased by 4.7% compared to 2021. We note that not all content actioned by community moderators may be harmful. Content may be violative of community rules; for example, not concerning the topic of the community. Reddit, 2022. Transparency Report. [accessed 4 March 2024]; Similarly, Nextdoor's 2022 Transparency Report shows that community moderators reviewed 92% of all reported content. Nextdoor 2022 Transparency Report. [accessed 4 March 2024].

[320] Mumsnet response to 2022 Illegal Harms Call for Evidence.; Nextdoor response to 2022Illegal Harms.; Wikimedia Foundation response to 2022 Illegal Harms Call for Evidence.

16.291 Some services already provide some materials to volunteer moderators. For example, Reddit provides a moderator help centre for its moderators, including courses for moderators.[321] Discord offers tools, resources and guidance as part of its 'Safety Library'.[322] Twitch provides various information pages on topics such as "Guide for Moderators", "Combating Targeted Attacks" and "Managing Harassment" to aid moderators. Wikipedia provides pages on standards requirements.[323]

16.292 We are therefore proposing recommending in our Illegal Content Codes of Practice that services which use volunteer moderation as a form of content moderation should ensure that volunteer moderators are provided with materials that enable them to fulfil their role. This should support volunteer moderators to be able to identify and take down illegal content swiftly in line with Measure 4A (have in place content moderation systems and processes) and 4B (internal policies).

Rights assessment

Freedom of expression and association

16.293 We would not expect this proposal to have specific adverse impacts on users' rights to freedom of expression and association. Volunteer moderators that are able to access material to enable them to carry out their roles should be likely to make more accurate decisions as they should be more aware of which content should be actioned and how. However, we acknowledge that services may potentially have less oversight and influence over volunteer moderators who are not employees or contracted staff because they may not have written contracts in place. We do not anticipate that the services this proposed measure applies to will rely on volunteer moderation alone due to the size of the service and degree of illegal content present on that service. It will also mean that volunteer moderators understand which content should be actioned, enabling some consistency in decision making. This will mean that decisions are likely to result in a more proportionate approach to content moderation by the service, and therefore likely to safeguard users' rights.

Privacy

16.294 We would not expect this proposal to have specific adverse impacts on users' rights to privacy. We consider that the provision of materials to moderators would help to safeguard users' privacy as they would have more information on which content should be actioned with support available from the service for any queries, resulting in more accurate decisions being made. However, we acknowledge that services may potentially have less oversight and influence over volunteer moderators who are not employees or contracted staff because

---

[321] Reddit provide a moderator help center containing links to the basics of stating a community on Reddit, overview and explanation of individual moderation tools, community engagement and advice and materials. They also provide sub reddits for news, support and requests. Reddit also provide volunteer moderators 'Reddit Mod Education Courses' . The layout and set up uses the way that the platform operates to provide materials and support to their users.

[322] Discord provide users information and links to support on its 'safety and moderation' page, including information on how to develop server rules, links to moderation and community support to manage your server and "handling difficult scenarios as an Admin". [accessed 24 April 2024].

[323] Twitch provide various pages that use pictures and videos to show the moderator's view of the channel and its tools.; Wikipedia's moderator access is dependent on hierarchy and are required to follow extensive procedural rule. Wikimedia, 2020. How Content Moderation and Anti-Vandalism Works on Wikipedia. [accessed 4 March 2024]; Oz, A., 2009. '"Move along now, nothing to see here": The private discussion spheres of Wikipedia', *SSRN*.. [accessed 4 March 2024].

they may not have written contracts in place. We do not anticipate that the services this proposed measure applies to will rely on volunteer moderation alone due to the size of the service and the degree of risk illegal content is present on that service. All services should ensure they comply with data protection laws and consider relevant guidance from the ICO.[324] Such compliance should reduce the likelihood of inaccurate personal data and ensure fairness in the processing of that personal data. It will also mean that volunteer moderators understand which content should be actioned, enabling some consistency in decision making. We consider this is likely to result in a more proportionate approach to content moderation and therefore tend to safeguard users' rights.

Impacts on services

16.295 Service providers are expected to incur direct costs if they need to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

16.296 For a service provider to implement this measure, there would be an initial cost of creating or providing the materials. This could be done either internally if they have the relevant expertise, or externally. Where a service chooses to source these materials externally, the cost of this would depend on whether the service already employs external organisations to provide materials for human moderators.

16.297 We expect that the majority of services within the scope of this measure would also be within the scope of Measure 4F (provision of training for paid moderators) in the Illegal Content Codes and could therefore build on or adapt the materials developed for this measure. There may be small additional costs associated with this e.g. adapting the format of the materials so that they can be accessed online rather than in person, making the materials searchable, or adjusting the level of detail so that the materials are relevant for the role of a volunteer moderator on that particular service. However, we do not anticipate that these costs are likely to exceed a few thousand pounds.

16.298 For services that rely only on volunteer moderators and do not have relevant existing materials developed for paid moderators, costs will be higher. However, we consider that the costs for these services associated with this measure would be considerably less than the cost estimates relating to providing training for paid content moderators that we outline in relation to Measure 4F in our Illegal Harms Consultation as these figures also include wage costs for moderators while receiving the training, which would not be incurred for volunteer moderators.

16.299 There would also be an ongoing cost to all services of updating the materials to ensure that they remain relevant. Even if it were possible to source an 'off the shelf' version of the materials, this would need to be updated and regularly reviewed in light of new and increasing kinds of illegal content.

16.300 The costs above are mitigated by the fact that this measure does not specify exactly how services should provide materials to volunteer moderators, giving services some flexibility to decide the most appropriate and proportionate approach for their own contexts.

[324] See ICO: Children's Code guidance and resources; A guide to the data protection principles and Content moderation and data protection. [accessed 4 March 2024].

16.301 In our Illegal Harms Consultation, we set out our provisional view that there would be a significant extra cost burden for services if we were to propose that all volunteer moderators be trained. We consider that the measure we are proposing now would represent a significantly lower cost burden to services, particularly given the flexible approach outlined.

Which providers we propose should implement this measure

16.302 As set out above, we consider that this measure could substantially reduce the risk of illegal content on U2U services by ensuring that volunteer moderators are better equipped to identify and swiftly take down illegal content where it is detected in these communities.

16.303 We expect that the costs for large services or multi-risk services of implementing this measure will generally be small, given that we expect the majority of these would also be within scope of Measure 4F (paid content moderation teams are appropriately trained) meaning that they will already have in place materials for paid moderators in most cases. The additional cost of making some of that material available to volunteer moderators should not be too substantial. The costs of this measure are also likely to be lower for smaller, less complex services with fewer risks, for example because their volunteer moderators are dealing with less complex issues and fewer kinds of illegal content.

16.304 For any services that rely only on volunteer moderators and do not have existing resources developed for paid moderators that can be adapted, costs will be higher. However, we consider the benefits from implementing the measure will also be higher, as having well-informed and well-prepared volunteer moderators will be particularly important where services place greater reliance on these moderators, to reduce the risk of moderation failures.

16.305 For any smaller services that are not multi-risk and use volunteer moderators, they would still be expected to consider how to provide relevant information to volunteer moderators if appropriate as part of implementing Measure 4A. However, we are not necessarily recommending they follow the specific approach to providing materials as described in this proposed measure. Benefits from the measure would be lower on these services, given limited risks of illegal content. The costs of this measure would also be substantial for such services, as they would not be in scope of Measure 4A and therefore may not have existing materials that can be adapted. Therefore, at this time we do not consider that it would be proportionate to recommend this measure for such services.

16.306 We are therefore proposing that this measure should apply to all U2U services that use volunteer moderation and are multi-risk (regardless of size) and all large U2U services that use volunteer moderation (regardless of risk level).

Provisional conclusion

16.307 Given the harms this measure seeks to mitigate in respect of illegal content, we consider this measure appropriate and proportionate to recommend for inclusion in the Illegal Content Codes.[325] For the draft legal text for this measure, please see 4G in Annex A9.

[325] This includes our Codes for Terrorism, for CSEA and for other duties relating to illegal content and harms.

17. Search moderation

This section outlines our recommendations for search services' moderation of their results to help them meet their children's safety duties. Content moderation entails U2U services reviewing content to decide whether it is content harmful to children and actioning identified content appropriately so as to protect or prevent children from encountering it. As it relates to search services, we refer to this practice as "search moderation" throughout this section.

In the Act, search services likely to be accessed by children are required to take proportionate steps to minimise the risk of children encountering PPC, PC and NDC. Search services can act as a pathway to harm by providing users, including children, access to content that may be harmful. We believe that effective search moderation plays an important role in protecting children from harm associated with PPC, PC and NDC; effective search moderation systems and processes will allow search services to identify and appropriately action content that is harmful to children.

We propose to adopt an approach consistent with that outlined in our previous 2023 Illegal Harms Consultation. The measures to protect children that we propose in this section should be considered separately, and in addition, to those outlined in the Illegal Harms Consultation. That is because there are differences in the duties underlying these measures that are specific to protecting children from harm. In this consultation we are proposing an additional measure related to services' safe search settings (see Measure SM2) that we believe will help providers of large general search services likely to be accessed by children meet the children's safety duties.

Our proposals

#	Proposed measure	Who should implement this[326]
SM1	Have moderation systems and processes in place to take appropriate action on PPC, PC and NDC A) When **PPC has been identified, downrank and/or blur the search content** B) When **PC and NDC has been identified, decide whether to downrank and/or blur the search content**	All search services
SM2	**When a user is believed to be a child, filter identified PPC out of their search results** through a safe search setting. Users believed to be a child should not be able to turn this setting off	All large general search services
SM3	**Set and record internal content policies**	All large general search services & All multi-risk for content harmful to children search services
SM4	Set **performance targets for search moderation functions**	
SM5	Develop and apply **policies on prioritisation of content** for review	
SM6	**Ensure search moderation functions are sufficiently resourced**	
SM7	Ensure people working in search moderation **receive training and materials**	

[326] These proposed measures relate to services likely to be accessed by children.

Consultation questions

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

39. Are there additional steps that services take to protect children from the harms set out in the Act? If so, how effective are they?

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?

The use of Generative AI (GenAI), Overview of Codes, Section 13 in this volume, to facilitate search is an emerging development, which may include where search services have integrated GenAI into their functionalities, as well as where standalone GenAI services perform search functions. There is currently limited evidence on how the use of GenAI in search services may affect the implementation of the safety measures as set out in this code. We welcome further evidence from stakeholders on the following questions:

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? Please provide arguments and evidence to support your view.

42. What additional search moderation measures might be applicable where GenAI performs or is integrated into search functions? Please provide arguments and evidence to support your view.

Introduction to search services and risks to children

Search services relevant context

16.308 Volume 3, Section 7.10, Risk of harm to children on search services explains that search services play a key role in making online content accessible to users, including children, and in shaping their online journeys. Our research has found that more than nine in ten (94%) children aged 8-17 claimed to use search services.[327]

16.309 Search services differ from U2U services in that they do not host content or facilitate interactions between users. Instead, search services help users access web pages and web hosted content by presenting them with search results. The content that appears in search results depends on the services' underlying search index and their ranking algorithms.[328] Search services often offer additional functionalities such as image or video search where the user is presented with relevant results in different formats that depart from the traditional text base "blue links" that directly connect to external websites. To facilitate search journeys, some services offer summary boxes providing a high-level summary response to the users' requests and have recently started to integrate GenAI to power responses to the users' questions. In linking users to relevant information, search services can act as a pathway to harm and present a risk of exposing children to content that is harmful to children, including PPC, PC and NDC.

---

[327] Ofcom, 2023, Children and Parents: Media Use and Attitudes. To clarify what was meant by search engine for respondents, children aged 8-17 were asked whether they used sites or apps like Google, Bing or Yahoo to look for things online. [accessed 6 January 2024].

[328] See Section 7.10, Risk of harm to children on search services for information on how indexing works on search services.

17.1 As set out in Volume 3, Section 7.10, Risk of harm to children on search services, we distinguish between the following types of search services: general and vertical search services.

17.2 General search services enable users to search the web by inputting search requests on any topic. Data from Ipsos Iris shows that Google Search and Microsoft Bing are the highest reaching search engines among UK online adults.[329] Ofcom's children's online passive measurement pilot study indicated that Google Search reaches 87% of online children aged 8 – 12.[330]

17.3 Vertical search services differ from general search services in that they only present users with results from selected websites with which they have a contract and an API, or equivalent technology is used to return the relevant content to users.

Search services and the risk of harm to children

17.4 Volume 3, Section 7.10, Risk of harm to children on search services and Volume 4, Section 12, Children's risk assessment guidance and risk profiles detail our understanding that large general search services and multi-risk services (which could include smaller general search services and vertical search services)[331] pose the greatest risks to children encountering harmful PPC, PC and NDC. In general, unless specialising in kinds of content that is harmful to children, vertical search services are considered to have an inherently lower risk given the far narrower scope of content presented to users that comes from pre-determined, often professional, or curated, locations on the web.

17.5 The main underlying risk of harm to children using search services stems from the users' ability to enter search requests related to PPC, PC and NDC and receive content that is harmful in the results. Effective moderation systems and processes are required to minimise children's risk of exposure to harmful content. As referenced in Volume 3, Section 7.10, Risk of harm to children on search services and Section 16, Content moderation for U2U services, ineffective or poorly resourced content moderation appears to have serious impacts on user safety on U2U services across a wide range of harms. Evidence suggests that service providers can increase user safety and reduce children's exposure to harmful content if they invest in improving content moderation systems. We assess that most evidence suggesting ineffective content moderation functions pose an increased risk of harm to users on U2U services, can be applied to search services.

Search services and user base

17.6 Our understanding is that some search services may employ technologies to profile users, including their age, based on user engagement and interaction with the search service. We, however, have limited evidence as to how they may do this. We understand that some search services allow users to self-declare their age on sign-up or when creating an account.

---

[329] Ipsos, Ipsos iris Online Audience Measurement Service, May 2023, age: 18+, UK. Google Search and Microsoft Bing reached 86% and 46% of UK online adults in May 2023 respectively. Note: Google Search does not include Google Search services - Maps, Shopping, Play or News. As reported in Ofcom, 2023. Online Nation.

[330] Ofcom Ipsos Children's Online Passive Measurement 2023, age: 8-12, UK. Base: 162. Data is not weighted. Due to low base size data should be treated as indicative only and not representative. As reported in Ofcom, 2023. Online Nation.

[331] See Section 7.10, Risk of harm to children on search services.

Exclusively relying on this information to target protections and experiences to children will put many children at risk as we know that children may input an incorrect age when opening online accounts.[332] For this reason, the Act is clear that self-declaration cannot be regarded as an age assurance method.[333]

17.7 We note that for U2U services, the Act requires certain services to use highly effective age assurance to prevent children from accessing PPC. The Act, however, does not require search services to use age assurance technologies to comply with the children's safety duties. Though it is something we may consider in the future, at this stage, we do not consider it proportionate to recommend that search services implement any form of age assurance to directly target their moderation actions to child users (as explained below in 'Other options considered' for Measure SM1). Instead, to ensure that children are adequately protected from PPC, PC, and NDC content, our proposed Measure SM1 may result in actions that impact both adult and child user access to content presented in or via search results. These impacts have been considered and balanced against the risk of harm to children.

17.8 If a service provider, however, does choose to use highly effective age assurance to target child protection measures exclusively to child users, then they may do so.

What are the duties in the Act?

17.9 The Act requires search services to take proportionate steps to:

a) mitigate and manage the risk of harm to children of different age groups (as identified in the services' risk assessment); and

b) mitigate the impact of harm to children of different age groups presented by search content that is harmful to children.[334]

17.10 The Act also requires that search services should operate using proportionate systems and processes designed to:

c) minimise the risk of children of any age encountering search content that is PPC; and

d) minimise the risk of children in age groups judged to be at risk of harm from other content that is harmful to children (including PC and NDC) encountering that search content.[335]

17.11 In practice, this means that search services are expected to minimise the risk of children encountering content that is harmful to them, via search results, by moderating search content on its service. It is important to recognise that in the Act, content is to be treated as 'encountered via' search results where it is encountered as a consequence of interacting with search results (for example by clicking on it).[336] The Act also states that this does not include a reference to encountering content as a result of subsequent interactions with an internet service other than the search service (for example, further clicks or interactions

[332] Ofcom, 2024. Children's Online 'User Ages'. Note: this study is about the user ages of children on user-to-user services, rather than Search, but the findings should have some applicability to Search services as well.
[333] Section 230(4) of the Act.
[334] Section 29(2) of the Act.
[335] Section 29(3) of the Act.
[336] Section 57(5)(a) of the Act.

with the site that the search result URL is linked to).[337] We, therefore, understand search content to include content on a webpage that can be accessed by users directly interacting with search results. The safety duties, and our recommended measures set out below, should be considered in this context.

17.12 The children's safety duties for search services (like the safety duties relating to illegal content) do not require search services to use 'content moderation, including taking down content' measures, but they do require search services to take or use 'content prioritisation' measures.[338] Additionally, search services have a duty to allow users to make complaints about content that is harmful to children and to take 'appropriate action' in response to such complaints.[339] Service providers may also take or use measures that result in content no longer appearing, or being given a lower priority, in search results.[340]

17.13 Search service providers therefore need a moderation function that enables them to make judgements about whether search content should be treated as content that is harmful to children, and to take appropriate action against that content to minimise the risk of children encountering it.

What is content moderation in search?

17.14 Section 16, Content moderation for U2U services acknowledges that content moderation systems and processes differ from service to service. We understand that services generally employ human review and/or automated technologies to identify and moderate harmful content, including illegal content and legal content that does not comply with its own content policies (i.e., violative content).[341] As acknowledged in our 2023 Illegal Harms Consultation and Section 16, Content moderation for U2U services most services detail what type of content is prohibited on their service in their public facing terms of service, which will normally comply with existing laws in different jurisdictions.

17.15 We recognise that there are different moderation actions that a search service might choose to apply for the purposes of complying with its duties. Actions may include downranking, blurring, filtering, or other forms of altering the prioritisation and visibility of content in search results. For the purposes of this section, references to 'search moderation' (and associated expressions) should be understood as referring to all such actions.

[337] Section 57(5)(b) of the Act.
[338] See, by comparison, sections 12(8)(e) and 29(4) of the Act.
[339] Section 32(2) and (5) of the Act.
[340] Section 32(5)(C) of the Act.
[341] Google states "Google's automated systems help protect against objectionable material. Search results should be useful and relevant, and limit spam responses. We may manually remove content that goes against Google's content policies, after a case-by-case review from our trained experts. We may also demote sites, such as when we find a high volume of policy content violations within a site." Google. Content policies for Google Search. [accessed 20 December 2023], Microsoft, 2023. Bing EU Digital Services Act Transparency Report. [accessed 20 December 2023]

Table 17.1: What are the different actions services might take to moderate search content?

Downranking: Involves altering the ranking algorithm to ensure that a particular piece of content appears lower in the search results and is, therefore, less discoverable to users.

Blurring: Involves obscuring the view of image-based content. For example, this may be done by a greyscale overlaying the image, accompanied by a content warning.

Filtering: Involves ensuring that content is not returned in search results based on whether a condition is/isn't met. For example, 'not displaying search results where condition "PPC" is true.'

Deindexing: Involves the removal of URLs (i.e., links to individual webpages) or domains (i.e. entire websites) from a search index. This will prevent the webpage URLs from appearing in search results entirely.

Delisting: Involves adding content to a blacklist to ensure it does not appear in the pool of content returned in search results. Content which has been delisted will still be found in the index.

17.16 Following feedback from our 2023 Illegal Harms Consultation, we have clarified that deindexing and delisting are separate actions and we do not treat them as interchangeable. We will also be considering replicating this clarification in our Illegal Content Codes ahead of the Illegal Harms Statement.

Interaction with Illegal Harms

17.17 In our 2023 Illegal Harms Consultation, we proposed the following measures regarding content moderation for Search services to be included in our draft Illegal Content Codes:

a) **Measure 1:** Have systems or processes designed to deindex or downrank illegal content of which it is aware, that may appear in search results.
b) **Measure 2**: Set and record internal content policies having regard to the findings of risk assessment and any evidence of emerging harms on the service.
c) **Measure 3**: Set and record performance targets for its search moderation function and measure and monitor its performance against these targets.
d) **Measure 4:** Prepare and apply a policy about the prioritisation of content for review.
e) **Measure 5**: Resource its search moderation function so as to give effect to their internal content policies and performance targets.
f) **Measure 6**: Ensure people working in search moderation receive training and materials that enable them to moderate content effectively.

17.18 See Section 12 of the 2023 Illegal Harms Consultation for a detailed discussion of the evidence and impacts of those measures.

17.19 We provisionally consider that all measures in the draft Illegal Content Codes are also proportionate for providers of a service likely to be accessed by children. As with the draft Illegal Content Codes, and in line with content moderation proposals for U2U services, we considered different approaches for these measures regarding whether to specify a) detail for how services should configure content moderation systems and process, b) the outcomes systems and processes should achieve, or c) factors services should regard in designing these systems and processes.

17.20 Our provisional view remains that option c) is the most proportionate approach as it raises standards whilst also allowing for flexibility given that there is no 'one-size-fits-all' approach

to content moderation across the sector. We set out below our detailed assessments of the evidence and impact of these measures as they relate to duties for services likely to be accessed by children.

17.21 We are also proposing to include an additional Measure (SM2) into the Children's Safety Codes of Practice that recommends the application of safe search where a large general search service believes a user is a child. As explained in our 2023 Illegal Harms Consultation, we view safe search largely as a tool that is most appropriate for controlling the search content that children might encounter as a means of complying with the children's safety duties.[342]

Our proposals to protect children

17.22 Given existing available evidence of risk and our understanding of current practice, we propose:

a) **Measure SM1:** All search services should have moderation systems and processes in place to take appropriate action on content that is harmful to children, which includes PPC, PC and NDC.
b) **Measure SM1A:** Service providers should downrank and/or blur all identified PPC.
   i) Providers should have regard to specified relevant factors to determine the extent to which they downrank and/or blur content.
c) **Measure SM1B**: Service providers should decide whether to take action on identified PC and NDC. If the provider decides to take action on identified PC and NDC, they should downrank and/or blur the PC and NDC content. Providers should have regard to specified relevant factors to determine:
   ii) If action should be taken on identified PC and NDC; and
   iii) If the provider decides to take action on identified PC and NDC, the extent to which they should downrank and/ or blur the content.
d) **Measure SM2:** In addition to the actions in Measure SM1A and Measure SM1B, large general search services should apply a safe search setting for all users believed to be a child which filters out identified PPC from search results. Users believed to be a child should not be able to switch this setting off.

17.23 We also propose that large general search services and search services that are multi-risk for content harmful to children:

e) **Measure SM3:** Set and record internal content policies;
f) **Measure SM4:** Set performance targets for its search moderation functions;
g) **Measure SM5:** Develop and apply policies on the prioritisation of content for review;
h) **Measure SM6:** Resource their search moderation function sufficiently; and,
i) **Measure SM7:** Ensure people working on search moderation receive training and materials.

17.24 We consider our measures to be a clear minimum basis for service providers to meet the duties as set out in the Act. We recognise that some providers, particularly large general search services, may already have in place some of the proposed moderation systems and

[342] See our Illegal Harms Consultation, Volume 4, Section 13.

processes. We are also aware that some providers may have in place moderation practices that go beyond our proposals. We encourage services to continue existing practice that may exceed our recommendations to protect children from harmful content in line with the requirements in the Act.

17.25 Our measures do not specify how service providers should identify content. We are aware that services may use a variety of tools, including reporting mechanisms, human reviewers, and automated systems, to identify content that is harmful to children and illegal content. We believe services are best positioned to determine the appropriate means to identify content. Thus, we propose to provide services the flexibility to implement the appropriate measures to meet their duties in a way that is cost-effective and proportionate for their circumstances, and that is consistent with ensuring that the risk of children encountering harmful content is minimised.

Relationship between publicly available statements and moderating content harmful to children

17.26 The Act requires that search services include provisions in their publicly available statements[343] specifying how children are to be protected from search content that is PPC, PC and NDC that is harmful to children.[344]

17.27 These duties apply across all areas of a search service, including the way the search engine is designed, operated, and used, as well search content that is allowed/not allowed on the service. When identifying kinds of content and making content judgements, providers have a choice: they may either use the categories of content defined in their publicly available statement, which should be at least as broad as those defined in the Act, or they should use the categories defined in the Act.

Measure SM1: All services should have moderation systems and processes in place to take appropriate action on content that is harmful to children.

Explanation of the measure and appropriate actions

17.28 In line with the safety duties in the Act, we propose that all search services likely to be accessed by children have moderation systems and processes in place to take appropriate action on PPC, PC, and NDC. We propose different approaches to moderating PPC compared to PC and NDC; we explain our rationale below.

17.29 Both Measure SM1A and Measure SM1B should be applied to all users unless a user is believed to be an adult based on reasonable grounds.

17.30 Measure SM1A: Service providers should downrank and/or blur all identified PPC

a) Providers should have regard to the below relevant factors to determine the extent to which they downrank and/or blur identified PPC.

[343] See Section 19, Terms of service and publicly available statements for more detail on duties pertaining to publicly available statements.
[344] Section 29(5) of the Act.

17.31 Measure SM1B: Service providers should decide whether to take action on identified PC and NDC.

17.32 If the provider decides to take action on identified PC and NDC, they should downrank and/or blur the PC and NDC content. Providers should have regard to the relevant factors below to determine:

a) If action should be taken on identified PC and NDC, and;
b) If the provider decides to take action on identified PC and NDC, the extent to which they should downrank and/or blur the content.

Factors relevant for Measures SM1A and SM1B

- **The prevalence** of PPC, PC and NDC hosted by the person responsible for the website or database concerned;
- **The severity of harmfulness** of the identified PPC, PC and NDC; and
- **The interests of all users** (**including children, but particularly adult users)** in receiving any content that is not harmful to children that would be affected by the action taken.

How to use the relevant factors for Measure SM1A

17.33 The factors listed above should help service providers determine the extent to which they downrank and/or blur identified PPC. We recognise that, depending on the circumstances, it may be appropriate for services to action content in various degrees by, for example, downranking by 100 places versus 20 places, or opting for a combination of downranking and blurring as opposed to just blurring.

How to use the relevant factors for Measure SM1B

17.34 The factors listed above will help service providers determine whether to take action on identified PC or NDC (i.e. when it is appropriate to downrank and/or blur content).

17.35 Where services decide to downrank and/or blur identified PC or NDC, the factors should also help them determine the extent to which they downrank and/or blur (as in Measure SM1A).

Who our measures apply to

17.36 We recognise that adults have a right to access content harmful to children that is captured by our proposed measures. Action taken by providers under Measure SM1A and Measure SM1B should therefore apply to all users, apart from users believed to be an adult based on reasonable grounds.

17.37 Reasonable grounds should be based on information that service providers' have or infer about a user's age – this could include, but is not limited to, means of highly effective age assurance. For example, where a user's age has been age assured by highly effective age assurance on a U2U service, which is provided by the search service provider, we would consider the service provider to have reasonable grounds to believe the user is an adult.

17.38 We take the view that where a service provider does not believe a user to be a child, this will not suffice as reasonable grounds that a user is believed to be an adult. Without reasonable grounds to believe a user is an adult, we consider that it is possible that the user could be a child and, therefore, Measures SM1A and SM1B should apply. For the avoidance of doubt, the measures will not apply to users believed to be adults on reasonable grounds, but will apply to all other users, which may include users believed to be children and users where the provider cannot confirm if a user is an adult or a child.

17.39 We note that Measure SM1B differs to Measure SM1A, because search services will be able to determine whether or not they take action on identified PC and NDC using relevant factors. However, we expect that where there are reasonable grounds to believe a user is an adult, both Measure SM1A and SM1B should not apply. This is because under SM1B, the 'relevant factors' considered by the service provider do not include consideration of whether a user is an adult or the number of users that are an adult. [345]

17.40 For the avoidance of doubt, we recommend that Measure SM1 (including the recommended appropriate actions in Measures SM1A and SM1B) apply to all search services likely to be accessed by children, including general search services of all sizes and vertical search services to the extent that content harmful to children may be encountered by children on or via those services. Please see 'Which providers we recommend implement this measure' below for more detail.

17.41 We would expect providers of large general search services to implement both Measure SM1A and Measure SM2 (which we impact assess separately below) when it identifies PPC, as well as Measure SM1B when it identifies PC or NDC.[346]

Downranking and blurring as appropriate actions to moderate content

17.42 We assess that if services downrank and/or blur search content that has been identified as PPC, PC and NDC, children may be diverted from search pathways that could result in a potential risk of harm. As such, we believe that services can minimise the risk of children encountering harmful content if they have in place the moderation systems and processes to take appropriate action on identified PPC, PC and NDC. As explained below, we consider that downranking and blurring can be appropriate moderation actions.

Actions may be applied via existing 'Safe Search' functionalities

17.43 As acknowledged in 'Current practice' for Measure 2 below, many general search services of different sizes have established a safe search function within which exist different settings offering different levels of visibility and access to certain kinds of content.[347] It is possible that some services may choose to implement the appropriate actions recommended in Measure SM1A and SM1B through their existing safe search function and default safe-search settings. We understand that the current practice among many search services of all sizes is to offer users the ability to switch off or change their default safe search settings, apart from when users are believed to be a child (see Current practice section in Measure SM2)/

17.44 If a service chooses to implement Measure SM1A and Measure SM1B through their existing safe search function, we consider that they would still be in the safe harbour even if users are able to switch off or change their default safe search settings such that the actions of downranking or blurring of PPC, PC, and NDC content no longer occur. However, for services in scope of Measure SM2, i.e. large general search services, we are proposing that users should not be able to switch this setting off.

17.45 We recognise that allowing users (which potentially could include children) to turn off their safe search settings to see PPC, PC and NDC that has been downranked or blurred by the service may impact the effectiveness of the measures. However, we have had regard to

[345] This example is in relation to the factors relevant to Measures 1A and 1B: Interest of all users (including children, but specifically adults).
[346] In circumstances where SM2 doesn't apply (all users, apart from users believed to be a child) then SM1A will apply.
[347] See 'Effectiveness at addressing risks to children', Safe Search settings.

evidence which shows that default settings are effective as users often do not change or move away from the default setting.[348] Following this evidence, we presume that children will be less likely than adults to move away from the default setting if they are not actively searching for specific kinds of content. In practice, this means that our measures should, in most cases, still effectively minimise the risk of children encountering harmful content, while minimising the impacts on adult users who would still wish to search for content which might include PPC, PC or NDC.

17.46 In the context of large general search services which present the greatest risk to children, we believe that our recommendation in Measure SM2 mitigates the risks associated with the above scenario; Measure SM2 specifies that PPC should be filtered for all users believed to be a child without the ability to switch off this setting.

Different appropriate action for PPC, PC and NDC

17.47 Our proposal to recommend a different approach for PPC on the one hand, and PC and NDC on the other, is grounded in the evidence we have to date with respect to the risk factors around search services and the extent of harm for each content type.

17.48 Our evidence in Section 7.10, Risk of harm to children on search services shows that some search functionalities are a particularly effective way for users to find some kinds of content, including PPC. This increases the risk of children encountering PPC when using such functionalities. For example, there is evidence that image search results may be more likely than text/URL results to contain content promoting self-injurious behaviour, eating disorders (which may be particularly risky when presented outside of their original context) and pornography.[349] Additionally, search request inputs have been found to be an effective vehicle for users who employ 'coded language', which is associated with harmful content to effectively access such content.[350] Further to this, evidence suggests that search services provide access to content that may be harmful to children with minimal friction, particularly if users are actively searching for such content. This content can appear high up in returned search results or ranking, increasing the likelihood of users, particularly children, being exposed to harmful content.[351] In particular, evidence suggests that among 16–21-year-olds who have previously seen pornography online, 30% reported having done so through search engines; 79% in this group had encountered violent pornography before the age of 18.[352]

17.49 There is further evidence in Section 7.10, Risk of harm to children on search services, demonstrating the extent of PPC-related harm experienced by children on, or via, search services. Evidence demonstrates that encountering suicide and self-harm content can exacerbate poor mental health in children and increase the risk of self-harm behaviours and suicidal ideation. In extreme cases, evidence suggests it may lead to children taking their own lives. A report which looked at deaths by suicide of children and young adults aged 10-19 in the UK (based on national mortality data between 2014-2016) found that almost a quarter (24%) of these children and young adults were known to have had 'suicide-related online experiences' (including actions such as searching the internet for information on

[348] Competition & Markets Authority, 2022. Online Choice Architecture, how digital design can harm competition and consumers. [accessed 3 December 2023].
[349] See Section 7.10, Risk of harm to children on search services.
[350] See Section 7.10, Risk of harm to children on search services.
[351] Ofcom, 2023. One Click Away: A Study on the Prevalence of Non-Suicidal Self Injury, Suicide, and Eating Disorder Content Accessible by Search Engines. [accessed 3 December 2023].
[352] Children's Commissioner, 2023. 'A lot of it is actually just abuse'. [accessed 6 November 2023].

suicide methods, visiting websites that may have encouraged suicide and communicating suicidal ideas online).[353]

17.50 For these reasons, we therefore propose to recommend that all PPC identified by the provider should be downranked and/or blurred for all users (unless the provider has reasonable grounds to believe a user is an adult, while allowing them some flexibility as to what extent services downrank and/or blur. In the case of PPC on large general search services, please also see Measure SM2.

17.51 However, there is more limited evidence of children's experience of PC-related harm on search services. Therefore, in relation to PC and NDC (Measure SM1B), we consider it would be appropriate to give services discretion and help services to decide when to take action (by downranking and/or blurring) on PC and NDC, as well as to what extent services downranking and/or blur when they decide to take action.

17.52 We acknowledge that there can be overlaps between some categories of PPC and PC, where there may be similarities in impact and accessibility in certain contexts. Therefore, we invite views and evidence from stakeholders to demonstrate if the approach towards PPC should be extended to PC and NDC.

Current practice

17.53 We understand that Google Search and Microsoft Bing, both large general search services who have control over their index,[354] use a combination of automated systems and human reviewers in their moderation functions.

17.54 Google Search primarily use automation to moderate policy-violating content. Google Search also relies on quality raters to assess the quality and usefulness of search results based on a variety of signals so as to improve its automation and inform search ranking.[355]

17.55 Microsoft Bing ranks search results based on relevance, quality, user engagement, freshness, location, language and page load time. Microsoft Bing may moderate search requests that could unexpectedly expose users to self-harm, violent, graphic or hateful content, or misinformation.[356] AI-based classifiers are used on search prompts and may lead to moderation actions (i.e. not returning generated content to the user or diverting the user to a different topic).[357] Microsoft Bing tracks accuracy metrics to monitor moderation effectiveness.[358]

17.56 We do not have the same evidence base for smaller services. Some smaller services may not be in control of search engine operations. In those cases, we anticipate that the search engine would be moderated by the upstream provider.

---

[353] However, the authors flag limitations: that this may be an under-estimate, as suicide-related internet use is not always documented, and causal links cannot always be identified. Source: Rodway, C., Tham, SG., Ibrahim, S., Turnbull, P., Kapur, N. and Appleby, L. 2022. Online harms? Suicide-related online experience: a UK-wide case series study of young people who die by suicide (p.4442). *Psychological Medicine*, 54 (4434-4445). [accessed 2 October 2023].

[354] See Section 7.10, Risk of harm to children on search services for information on how indexing works on search services.

[355] Google response to 2023 Protection of Children Call for Evidence. [accessed 6 November 2023].

[356] Microsoft Bing, no date. How Bing delivers search results. [accessed 10 October 2023].

[357] Microsoft Bing, 2023. Bing EU Digital Services Act Transparency Report. [accessed 10 October 2023].

[358] Microsoft Bing, 2023. Bing EU Digital Services Act Transparency Report. [accessed 10 October 2023].

17.57 We similarly have limited evidence related to the practices of vertical search services. However, given the nature of these services, which draw search results from pre-determined websites about specific topics and genres, it is unlikely that children will encounter PPC, PC and NDC. As such we believe vertical services will have fewer moderation systems in place to take action on such content. This, in turn, means there is less evidence we are able to draw into our current practice section.

Effectiveness of blurring and downranking to minimise the risk of harm (Measures SM1A and SM1B)

17.58 As outlined in Section 7.10, Risk of harm to children on search services, we know that children can be exposed to harmful content online by accessing content through search services. This includes content that encourages, promotes or provides instructions for primary priority harms such as suicide, self-harm, eating disorders [359] and pornography. [360] This content can:

a) appear in search results and be accessed by children;
b) rank highly (be found at the top) in search results, increasing the likelihood of children clicking on it and being exposed to harmful content; and
c) appear in visual form in image search, evoking a shocking emotional response.

17.59 There are a range of steps that a service may take to minimise the extent to which children encounter search content, from actions that ensure that the content is not included in search results for all users (i.e. deindexing or delisting) or some users (i.e. filtering), to those that reduce the ease of access or visibility of search content (ie. downranking or blurring).[361] We explain the range of actions used by services to moderate content in Table 17.1 above.

17.60 In Measure SM1A and Measure SM1B, we propose to recommend downranking and blurring as appropriate actions that can be applied to content that is harmful to children, because content which has been downranked or blurred will still be returned in search results and can still be accessed by all users. This means that while the risk of children encountering the content is materially reduced, it remains accessible to users who could be adults (as the service provider does not have reasonable grounds to believe they are adults) and these users will also be impacted by the application of these moderation actions.

17.61 We acknowledge that an argument might be made that downranking and blurring may not be sufficient as children will still be able to access the relevant content via the service (for example, by scrolling further down or clicking through to view blurred content at the URL on which it is hosted). While alternative actions that result in the content no longer appearing in search results may be more effective at eliminating the risk to children, we consider such a recommendation would not be justified on proportionality or freedom of expression grounds in circumstances where the moderation actions will be applied for all users of a service, and where the content is not illegal. Nonetheless, we consider that the actions of downranking and blurring will contribute to reducing the risk of children encountering and

[359] Ofcom, 2024. One Click Away: A Study on the Prevalence of Non-Suicidal Self Injury, Suicide, and Eating Disorder Content Accessible by Search Engines. [accessed 6 November 2023].
[360] Children's Commissioner, 2023. A Lot of it is Actually Just Abuse – Young People and Pornography. [accessed 14 September 2023].
[361] For definitions of these terms please see Table 17.1.

being harmed by said content, compared to the alternative where this content remains easily discoverable via search results. We explain these actions further below.

Downranking of search content

17.62 Downranking content involves altering the ranking algorithm to ensure that a particular piece of content appears lower in the search results and is therefore less discoverable to users. Downranking imposes a degree of friction to the user's search experience; users might have to spend longer scrolling down to access or view downranked content. imposing a degree of friction to the user's experience of search engines. In many cases, we understand that providers already downrank content that breaches their policies.

17.63 Evidence indicates that the first page[362] of search results is the most accessible to users.[363] These findings suggest that the first page on a search service is the most relevant when examining the content that most users will encounter. By extension, this suggests that when a service downranks content that is harmful to children, the content will be harder for users to find and the risk of a child encountering that content will be minimised, in line with the requirements of the children's safety duties. It follows that this content is less likely to cause children, and others, harm.

17.64 As explored in Section 7.10, Risk of harm to children on search services, content that encourages, promotes, or provides instructions for suicide, self-harm, and eating disorders can appear high up in returned search results or ranking, increasing the likelihood of children's exposure to harmful content. We therefore consider that downranking is an appropriate action services can take to meet their children's safety duties, as it can help to ensure harmful content is less easily accessible to children.

17.65 We expect pornographic content in search results to be downranked. This could include, for example, a pornographic image presented as a search result or a URL to a webpage on which pornographic content is accessible with one click on the URL. We note, however, that some of the requirements of the Act require providers of services that host pornographic content to use highly effective age assurance to prevent children from accessing pornographic content on the service. Where highly effective age assurance is in place, we anticipate that clicking on the search result would not present users with pornographic content. As noted in 'What are the duties in the Act' above, we consider 'encountering via' search results to be the consequence of direct interaction with search results (i.e. by clicking on it).[364] Therefore, where a URL search result connects to a service which hosts pornographic content, and the URL only leads to a webpage which requires an age check to be carried out, and which does not present pornographic content directly to the user, then the URL search result would not need to be downranked.

17.66 We considered recommending a set amount of places content should be downranked, for example further than the first page of results given cited evidence that the first page of search results is the most accessible to users.[365] Specifying the ranking position of search content raised technical and cost concerns about how services could effectively implement this recommendation. We also consider that the most appropriate downranked position is

[362] Note that some search services have replaced 'pages' with a continuous scroll like function.
[363] Beus, J, 2020. Why (almost) everything you knew about Google CTR is no longer valid. [accessed 15 April 2024].
[364] Section 57 (5) (a) of the Act.
[365] Beus, J,2020. Why (almost) everything you knew about Google CTR is no longer valid. [accessed 15 April 2024].

likely to depend on the factors that we recommend services regard, such as the nature and severity of the content.

Blurring of search content

17.67 Blurring involves obscuring the view of image-based content. Content which has been blurred will still be returned in search results and can be accessed by users by clicking through the attached link or interstitial. While we recognise that this may reduce the effectiveness of the action, we nonetheless consider that it provides an appropriate degree of friction for children and thereby helps services to meet their children's safety duties without unduly interfering with the rights of adults to receive the underlying content.

17.68 Section 7.10, Risk of harm to children on search services, cites research that image results surfaced in response to suicide, self-harm and eating disorder-related search requests presented a greater proportion of harmful content than other forms of search results. Further to this, upon exposure to images containing harmful content, users can experience strong, immediate, emotional responses.[366]

17.69 Blurring image-based search content can ensure that content is less immediately visible to users when viewing search results returned by a service. While the evidence and reasoning relate to PPC, it can be extended to PC or NDC visual in nature; blurring such visual content will both reduce the risk of children encountering it and minimise shock from the initial exposure to content, resulting in less harm to children.

17.70 Research on U2U services suggests that both adults and children see value in blurring because it physically prohibits interaction with certain kinds of disturbing content.[367] While this evidence relates to user experiences on U2U services, we believe this emphasizes the effectiveness of blurring image-based search results containing harmful content on search services. Blurring can help add friction to the search pathway and prevent the risk of harm to children. It may also benefit children who accidentally enter terms relating to harmful content.

17.71 We recognise that not all search services in scope of Measures SM1A and SM1B currently have a blurring function. While we are aware that Google Search blurs content,[368] we are not aware of the extent to which other services use blurring as a tool to moderate potentially harmful content. For example, our evidence suggests that Microsoft Bing currently does not blur content as a form of content moderation. However, the Microsoft Bing product Copilot AI appears to employ blurring of faces in images uploaded by users,[369] which suggests that Microsoft Bing can use blurring in some capacity for some processes.

17.72 While blurring may be particularly effective for, and tailored to, image-based results, we consider that downranking is just as effective as blurring at minimising the risk of children encountering harmful content. For this reason, we do not propose to prescribe blurring as the only action for image-based results and leave it to services' discretion to choose the most appropriate action. We understand that downranking content using existing infrastructure will likely be of lower cost to services than developing new infrastructure to blur content for some services.

---

[366] Ofcom, 2024. One Click Away: A Study on the Prevalence of Non-Suicidal Self Injury, Suicide, and Eating Disorder Content Accessible by Search Engines. [accessed 9 February 2024].
[367] Ofcom, 2023, User attitudes to On-Platform interventions. [accessed 9 February 2024].
[368] Google, no date. SafeSearch settings. [accessed 26 February 2024].
[369] Bing, no date. Copilot in Bing: Our approach to responsible AI. [accessed 26 February 2024].

17.73 We also recognise that blurring may be implemented in multiple ways. For example, some services may apply a content warning, also known as a sensitivity label, alongside the blurring. We considered whether it would be appropriate to recommend that services include content warnings alongside blurring. We understand that content warnings aim to inform/warn the user that they could encounter harmful content, and therefore may generally be useful tools. However, there is not enough evidence to support that content warnings would particularly benefit children and contribute further to reducing the risk of children encountering harmful content. For example, it is not unreasonable to imagine a scenario where some children may be more inclined to click through and reveal an image if they have been informed by the service that it may be somewhat harmful.

17.74 Services that do not currently apply content warnings may also incur additional costs to develop this technology, which we do not consider appropriate in the absence of evidence. However, we consider that at present services are generally best placed to determine how to blur content in a manner that is most effective. Where services may apply content warnings in addition to blurring, our proposed measure is not intended to be a signal that this would be disproportionate, and we would encourage services to continue to do so.

17.75 We note that service providers may have systems in place that enable them to take additional, more severe, action on content, such as deindexing, delisting or filtering. If a provider would rather use these actions to moderate content rather than our recommended actions of downranking and/or blurring, they may do so as long as they fulfil their children's safety duties.

Factors relevant to determination of appropriate action

17.76 The relevant factors (prevalence of PPC/PC/ NDC, severity of harmfulness and interest of all users) are adapted from those recommended for service providers in Measure SM1 of our Illegal Harm's Consultation, where they are used in reference to different recommended actions specific to the Illegal Harms context.[370] These actions include deindexing and delisting.

17.77 We consider that these factors are also relevant to the search moderation determinations required by Measure SM1A and Measure SM1B, as explained below.

a) **Measure SM1A:**
   i) to determine the extent to which identified PPC should be downranked and/or blurred.
b) **Measure SM1B:**
   i) to determine if action should be taken on identified PC and NDC, and
   ii) if services decide to take action on identified PC and NDC, to help them determine the extent to which such content should be downranked/blurred.

17.78 The factors are intended to enable services to weigh up the risks of harm from content against users' freedom of expression rights, based on the individual circumstances.

Prevalence of PPC/ PC/ NDC

17.79 Prevalence of PPC/ PC/ NDC hosted by the person responsible for the URL or database alongside content that is not content that is harmful to children.

[370] See our Illegal Harms Consultation, Volume 4, Section 13, Page 59.

17.80 Unlike U2U services, search services are unable to moderate individual pieces of content as they do not have control of the content that is hosted by the person responsible for the underlying URL or database. Where PPC/PC/NDC is hosted alongside other non-harmful content, moderation actions taken to reduce the risk of children encountering harmful content will inevitably impact that other content. Services should therefore consider the relative prevalence of content that is harmful to children when deciding which action, and the extent of such action, is appropriate.

17.81 Google Search's content policies state that Google demotes content when they find a "high-volume of policy content violations;"[371] we do not know what these high-volume thresholds are, but we recommend search services take into account the impact of high volumes of potentially harmful content when determining appropriate action related to search content that contains content that is likely to be harmful to children.

Severity of harmfulness

17.82 Severity of harmfulness, including whether the content is PPC/ PC/ NDC.

17.83 As outlined above, the Act distinguishes three different priority levels of content that are harmful to children (PPC, PC and NDC).[372] Within those categories of content harmful to children, there may also be a scale of seriousness; that is, some forms of the same content type may be more egregious than others based on the precise wording or presentation of the content (i.e. the difference between 'mild violence' and 'graphic violence'). Therefore, it is reasonable to expect providers to consider the severity of potential harm posed by search content that is PPC, PC and NDC in determining what action might be appropriate in respect of that content. We recommend service providers refer to our draft Guidance on Content Harmful to Children where we provide examples of PPC and PC.

17.84 The Act does not provide specific direction or language to determine what 'severity' or 'degrees of harm' look like for content that is harmful to children. We will not recommend how platforms should determine severity, acknowledging that platforms will have risk assessments and internal content moderation policies to draw from (see Measures 3 and 4) to assess what presents a significant material risk to children on their platforms, and can refer to our Register of Risk and Risk Profiles for additional guidance. We also acknowledge that there is limited evidence of existing practice. We therefore do not consider it would be appropriate to prescribe how severity analysis should be conducted.

Interest of all users (including children, but particularly adult users)

17.85 Interest of all users (including children, but particularly adult users) in receiving content which is not PPC/PC/NDC that would be affected. A service should, for example, consider the existence of other content that is not harmful to children that is present on a webpage and, therefore, the impact of any moderation action taken on that content.

17.86 We considered that it may be difficult to assess when action should be taken on a URL which leads to content that contains some PPC, PC and NDC, alongside content which may be recovery-focused and not harmful (in the case of suicide, self-harm or eating disorder content) or unrelated content that is not harmful to children. We therefore think it is

[371] Google, no date. Content policies for Google Search - Google Search Help [accessed 12 September 2023].
[372] Section 60(2)(c) of the Act defines non-designated content as content that is neither primary priority content or priority content which presents a material risk of significant harm to an appreciable number of children in the United Kingdom.

relevant for services to consider how moderation action taken on that content will impact all users (but particularly adult users, although also including children) and the person responsible for the relevant URL or database.

Identifying content harmful to children

17.87 We will not prescribe how search services should identify content for any of our search moderation measures. As referenced above, service providers can use categories of content defined in their publicly available statement (where these are broad enough to cover relevant forms of content harmful to children) or categories defined in the Act (see 'Relationship between publicly available statements and moderating content harmful to children'). We recommend service providers refer to our draft Guidance on Content Harmful to Children for examples of PPC and PC. It will be for providers to decide how to identify and label content as PPC/PC/NDC. We understand that providers may identify content through:

a) user reporting and complaints channels; or
b) the use of automated technologies, including existing technologies that underpin their safe search functionalities.

Other options considered

17.88 We considered whether it would be appropriate to recommend that the appropriate moderation actions specified in Measure SM1 be applied to child users specifically. We are not proposing this for two key reasons:

a) First, we do not have evidence to suggest search services are currently using forms of highly effective age assurance to identify child users. While services may give users the opportunity to make 'child accounts' and/or use signals about age to get an indication of which users are children, [373] there is insufficient evidence for us to deem these methods alone meet the criteria of highly effective age assurance that can determine which users are children on a service. Therefore, based on current practice, we do not consider it likely that search services would be able to robustly identify child users so as to ensure they benefit from the protections of this proposed measure exclusively.
b) Second, as noted above in 'Search services and user base' the Act does not require search services to use age assurance technologies to comply with the children's safety duties. We consider that it would be disproportionate to recommend the use of age assurance technologies in our search moderation measures given the nature of search services and how they operate; to require every user to create an account and undergo an age check to use a search service may have privacy and freedom of expression impacts on users, as well as risk fundamentally changing the business model of search services in comparison to U2U services.

[373] Where personal data is collected by the service provider to get an indication of a user's age, providers must be compliant with relevant data protection requirements for data minimisation and purpose limitation. See ICO, Expectations for age assurance and data protection compliance (Principles 6.1.4 Purpose limitation and 6.1.5 Data minimisation) for more information.

Rights assessment (Measure SM1A and SM1B)

17.89 As with content moderation by U2U services, see Section 16, search moderation is an area in which the steps taken by services as a consequence of the Act may have a potentially significant impact on the rights of users, in particular, their rights to privacy (Article 8), freedom of religion and belief (Article 9) and freedom of expression (Article 10). We have, therefore, considered the extent to which the degree of interference with these rights is proportionate.

17.90 As outlined earlier in this section, the moderation actions taken by services in line with our proposed Measures SM1A and SM1B will be applied to all users, apart from users believed to be an adult based on reasonable grounds, and therefore may interfere with the rights not only of children (the protection of whom the measure is designed to secure), but also of users who could be adults due to the service provider not having reasonable grounds to believe they are an adult. We consider those impacts below.

17.91 By limiting children's exposure to content that is harmful to them in this way, the proposed measure will seek to secure adequate protections for children from harm, in line with the legitimate aims of the Act. The moderation of content harmful to children acts to minimise the harmful consequences of such content, which can include harm to children's physical, mental or emotional wellbeing. We consider that a substantial public interest arises in relation to this proposed measure in the protection of children's health and morals, public safety and, in particular, the protection of the rights of others, namely child users of regulated services.

Freedom of expression

17.92 As explained in Volume 1, Section 2, Article 10 of the ECHR upholds the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. It is a qualified right, and Ofcom must exercise its duties under the Act in a way that does not restrict this right unless satisfied that is necessary and proportionate to do so.

17.93 We have carefully considered the impact of our measure on users' rights to freedom of expression, including the right of services to impart information, and users' right to receive information and ideas. We understand that our proposed Measures SM1A and Measure SM1B will impact the ease with which users' access PPC, PC and/or NDC and are mindful that this content is legal. We acknowledge that Measure SM1A in particular would require search services to limit the visibility and prominence of identified PPC for all users, including adult users, including for pornographic content, and that search services are a common way for adult users to choose to seek out this form of content. Therefore, we recognise that this measure could have a significant impact both on their rights to search for, and thereby access, such content, and on the rights of interested persons who make such content available to adult users. It would also, in a similar way, affect the rights of search service providers to make such information available to their users.

17.94 While we recognise the negative impacts that may result on the adult user experience (and therefore on the rights of interested persons and search services as well), our measure will not result in the removal of content from the search results, and under our proposed measure, adult users will still be able to search for, and access, the information, if desired. In addition, we consider the following aspects of the proposed measure would mitigate the impact on adult users' ability to search for such material:

a) The measure would only impact search content that is identified as PPC, PC and NDC. It would not therefore require search services to downrank or blur (or take any other appropriate action in respect of) search content which might surface content that is harmful to children elsewhere on a website, but which is not accessible through one click from the search results. This means, for example, that where services that host pornographic content are deploying highly effective age assurance to secure that children are not normally able to access pornographic content on the service (e.g. by putting in place an age check requirement on the domain page, and no pornographic content is visible prior to the completion of the age check), there would be no need to downrank any link to that webpage.
b) We acknowledge that one way in which search services may choose to implement this measure could be by way of a safe search feature, implemented as a default setting for all logged-out and logged-in users, but which users could choose to turn off (subject to the requirements of Measure SM2, which we propose to apply to large general search services only and would apply to users believed to be a child only). If this is the way that search services choose to implement this measure, then adult users would be able to disable this setting and obtain search results without downranking or blurring of PPC, PC and NDC if they choose. (The 'Actions may be applied via existing 'Safe Search' functionalities' section above explains why we consider that it is less likely that children would opt to do this and why we therefore think this measure would still provide them with adequate protection from PPC, PC and NDC in search results).
c) While we are not recommending the use of age assurance in support of this measure for the reasons set out above in 'Other options considered', where search services believe that some of their users are adults as a result of highly effective age assurance, then we have acknowledged that it would not be necessary for search services to apply these protections to those adult users. See 'Who our measures apply to' above for more detail.

17.95 We also consider that, while there is potential risk for a margin of error in search moderation, services have incentives to limit the amount of content that is wrongly actioned, to meet their users' expectations and to avoid the costs of dealing with appeals. Where a service decides to take action resulting in content no longer appearing in search results or being given a lower priority in order to comply with its children's safety duties, complaints procedures operated pursuant to section 32(2) of the Act should allow for the interested persons to complain and for appropriate action to be taken in response. The complaints process may also mitigate the impact on the interested persons' right to freedom of expression by giving them a mechanism for redress and providing a route to rectify any negative impact by having their content restored to an equivalent position to the one it would have been in had the action not been taken.[374]

17.96 Impacts on freedom of expression could, in principle, arise in relation to the most highly protected forms of speech, such as religious expression[375] or political expression, and in relation to the kinds of content that the Act seeks to protect, such as content of democratic importance and journalistic content. However, we consider there is unlikely to be a systematic effect on these kinds of content: for instance, such content would be unlikely to be particularly vulnerable to being wrongly classified as content harmful to children. In

[374] See Section 18, User reporting and complaints.
[375] Which could also engage users' or interested persons' rights to religion or belief under Article 9 of the ECHR.

addition, we have provided examples of the kinds of content, including more protected forms of speech, in our Guidance on Content Harmful to Children, which we encourage service providers to regard in implementing this measure.

17.97 For these reasons, we consider it unlikely that a less restrictive approach to search moderation could be adopted while still securing that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that the proposed measure is therefore proportionate.

Privacy

17.98 As explained in Volume 1, Section 2, Article 8 of the ECHR confers the right to respect for individuals' private and family life. We do not consider that moderation of search content in line with this proposed measure, whether by an automated or a human search moderation function, would amount to an interference with any user's rights to privacy under Article 8 ECHR. Search content identified as harmful to children and actioned through search moderation functions is, by definition, either identified in a way that enables a general search service to have made it available via search results or is made available for publication by a vertical search service under a bilateral contract with the content provider. This content would not, by its nature, contain information about any users of the service that requires processing in the identification of content harmful to children or application of an action, and the actions we recommend in our proposed measure would also not include any action against individual users.

17.99 We acknowledge that it is possible that the way search services decide to implement this measure could involve the processing of users' personal data and in this way may impact users' rights to privacy – for example, if search services decide to implement the measure in a way that gives users the option to turn off a safe search setting, or if search services choose not to apply these protections to users believed to be adults due to having highly effective age assurance information, and need to process their personal data to give effect to this. However, we are not specifying what forms of personal data they should gather to enforce their content polices and give effect to this measure, so long as they comply with the Act and the requirements of data protection legislation. We therefore consider that (assuming service providers also comply with data protection legislation requirements) the impact of the proposed measure as a result of services' search moderation decisions and processes on child and adult users' rights to privacy, above and beyond the requirements of the Act, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

Impacts on services (Measure SM1A and SM1B)

17.100 For these measures we have not set out how services should identify content, and we have therefore not quantified the cost of this process. In practice, we expect content may be identified through a service's reporting and complaints channel, or through automated labelling systems.

17.101 The costs of these measures will vary by service. For smaller services with low risks and few complaints about harmful search results for children, the costs should be minimal. Services with relatively few pieces of content harmful to children (for example, a vertical search service with a less extensive index of predominantly non-harmful content) should be able to assess and consider what action to take in relation to content when it is flagged or identified

for review. These services may not have to rely on automated systems to action harmful content, and instead may be able to action content on a case-by-case basis.

17.102 We consider there will be considerable costs to larger services and services posing significant risk to children. Such services may identify a high volume of suspected harmful content and the moderation systems and process to review this content may require substantial resources. As well as using human moderators, these services may have to employ automated systems to action and handle harmful content given the volume of content higher risk services encounter. This will entail higher costs to develop. For such services, we expect that these costs of assessing potentially harmful content and deciding how to action this content will account for most of the cost associated with this measure, but we have been unable to quantify this given that it is highly dependent on each service's approach, size and extent of risk.

17.103 In addition, services will incur costs associated with adapting their ranking algorithm to negatively weight content it has assessed needs actioning to reduce risk of harm to children. Additionally, services will incur a build cost when implementing a system to blur images or videos containing content harmful to children.

17.104 We believe that most of the cost to develop downranking and blurring systems will come from the quality assurance process. Services will have to test whether the interventions are applied accurately for content harmful to children and are not being applied inaccurately.

17.105 For a smaller search service, which does not already have downranking and blurring systems in place, we estimate it could take approximately 3-4 weeks of software engineering time, with an equivalent amount of non-engineering time, for each of these changes (downranking and blurring). Considering the labour costs presented in Annex 12, we expect the one-off direct implementation costs for each system could be around £7,000 to £18,000. We expect there will be additional incremental costs to maintain the systems and make sure that they are up to date. Assuming an annual maintenance cost of 25% of the implementation cost, this could be £2,000 - £4,500 per annum for each system; the cost of developing both systems would be double.

17.106 Costs may be higher for large services where we understand that significant review, coordination and governance processes may need to be followed to implement changes of this kind.

17.107 There would also be a cost associated with designing and implementing the ability for users to turn off their settings should a provider choose to implement this measure through their safe search function. We understand that these costs are likely to be minimal.

17.108 We recognise that many services may already have systems in place to moderate search content, such as the ability to blur images and videos, or downrank content. The costs to these services will depend on whether those systems are currently used to action content harmful to children. If services do not currently use these moderation systems to action content harmful to children, they will have to ensure that the systems are adapted to action content harmful to children. Services will also incur costs to test that the interventions are applied accurately as a part of their quality assurance process.

Which providers we propose should implement these measures (Measure SM1A and Measure SM1B)

17.109 We propose that these measures should apply to all search services likely to be accessed by children upon identifying PPC, PC and NDC. We consider these recommended measures are the minimum that services likely to be accessed by children should do to meet their safety duties in the Act and protect children from encountering PC and NDC.

17.110 We believe that the recommendations as presented in Measures SM1A and SM1B should assist services in meeting their children's safety duties. Namely moderation processes ensure services can identify and action content that is harmful to children, therefore minimising the risk of children of any age encountering PPC, and to minimising the risk of children in age groups judged to be at risk of harm from PC and NDC, from encountering such content.

17.111 As set out above, we believe the cost of taking appropriate action is likely to scale with a service's level of risk, and, therefore, also scale with the benefit of the measure. We consider this measure to be proportionate considering the various costs that some services may face, and the presented risk of harm to children absent moderation systems and processes.

17.112 We consider that there is a risk of particularly severe harm to children from encountering PPC on search services, especially via large general search services. We therefore believe that large general search services should go further in the case of PPC. In these instances, we propose to recommend Measure SM2 in addition to Measures SM1A and SM1B.

17.113 For large general search services and search services which are multi-risk for content harmful to children of all sizes (which may include vertical search services), we consider that these Measures SM1A and SM1B alone would be insufficient. Such services operate in a more complex risk environment, and therefore we consider it proportionate to further specify how they should design their policies, processes, frameworks and resources to moderate content effectively. The other proposed measures discussed in the rest of this section – SM3 to SM7 – consist of a package of further steps that we recommend such services should take.

17.114 We note that large general search services typically have characteristics that would make them likely to be multi-risk. These services tend to be widely used by children as well as adults, and they are designed to facilitate access to wide-ranging content, which may include large volumes of different kinds of harmful content.[376] However, throughout this section we still consider whether measures should also apply to large general search services that are not multi-risk, should any such services exist now or in the future.

17.115 For any smaller general search services which are not multi-risk for content harmful to children, and for vertical search services of all sizes which are not multi-risk for content harmful to children, we expect these measures SM1A and SM1B to provide adequate protection. Such services are less likely to face high volumes of diverse content that is potentially harmful to children that they need to assess. Given that these services operate in a simpler risk environment they could reasonably be expected to meet their child safety

[376] As set out in our children's risk assessment guidance, the outcome of service risk assessments is likely to depend on factors including the service's reach among children and the nature of content on the service. See Section 12 for more information.

duties without employing more sophisticated formal processes and frameworks. Where these services are operated by small or micro businesses with relatively limited resources, children may benefit more from resources being channelled toward core activities such as moderating content, rather than diverted towards additional, more complex systems and processes that may have only small incremental benefits on such services. In any case, smaller services that are not multi-risk should take all necessary steps to give effect to Measure SM1, even if we leave them more flexibility in how they approach this.

Provisional conclusion (Measure SM1A and Measure SM1B)

17.116 Given the harms this measure seeks to mitigate in respect PPC, as well as the risks of PC and NDC, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCS B1 in Annex A8.

Measure SM2: Large general search services should filter out PPC for users believed to be a child through safe search settings

Explanation of the measure

17.117 We propose to recommend that when the provider of a large general search service has identified PPC, it should:

a) apply a safe search setting for all users believed to be a child which filters out identified PPC from search results; and
b) take steps to ensure that this safe search setting cannot be switched off by the users believed to be a child.

17.118 We outline the current 'safe search' practices of large general search services below. We expect that in scope services will implement this measure through existing safe search settings. We expect that our proposed recommendation will involve expanding the scope of the kinds of content covered by existing practices to apply to all forms of PPC.

Who our measure applies to

17.119 This measure would apply in addition to Measure SM1A and Measure SM1B and the associated appropriate actions recommended for all users of the service, other than users the service has reasonable grounds to believe to be adults. As referenced above, large general search services pose a greater risk of harm to children compared to other services given their reach and ability for users to enter search requests related to any content type. We understand that large general search services have the technical capability and resources which allow us to recommend they take greater steps to protect children from PPC.

17.120 We propose to target this measure specifically at users believed to be a child. We refer to 'users believed to be a child' rather than 'child users' or 'children' to account for the efforts service providers have in place to profile users and to clarify that 'users believed to be a child' are not determined to be children through highly effective age assurance methods.

Filtering PPC for users believed to be a child

17.121 We propose to recommend that where a service believes a user to be a child based on indicators of age[377], a safe search setting should be applied which will filter out PPC from all search results. We believe this filtering will provide the safest search experience for children because it will ensure that PPC is not returned in search results.

17.122 Users that are believed to be a child should not be able to turn off their safe search setting to see unfiltered PPC. Providers should take steps to ensure the safe search setting cannot be turned off. Please see the section 'Effectiveness at addressing risks to children' for more detail.

Effectiveness at addressing risks to children

Safe search settings

17.123 Many general search services have so-called "safe search" features (safety settings) that reduce the discoverability of certain kinds of content. These safety settings are often applied by default for users and services generally provide a tool by which users can change these settings (either by switching them off or increasing/decreasing the level of protection).

17.124 As discussed, we expect that the large general search services in scope of this measure will implement the filtering of PPC through their existing safe search settings.

17.125 Please see the diagrams below which explain the different safe search settings for Google Search[378] (Figure 17.1) and Microsoft Bing[379] (Figure 17.2). The diagrams show the type of content covered in the service providers safe search settings (for example pornographic content), the action which is taken in each setting (for example blurring images) and the default settings applied to users (for example 'all users' and 'users with a child account').

17.126 Google Search and Microsoft Bing both operate a three-tiered approach to their safe search settings known as "Bing SafeSearch" and Google's "SafeSearch" feature. Each tier applies a different level of restriction on content for users. The highest safety setting (Tier 1) is applied by default to all users a service believes to be a child. The middle setting (Tier 2) is applied by default to all users. The lowest tier (Tier 3) turns off safety settings and associated restrictions; all possible search results are made available and surfaced to users.

[377] Where personal data is collected by the service provider to get an indication of a user's age, providers must be compliant with relevant data protection requirements for data minimisation and purpose limitation. See ICO, Expectations for age assurance and data protection compliance (Principles 6.1.4 Purpose limitation and 6.1.5 Data minimisation) for more information.

[378] Google, no date. Filter or blur explicit results with SafeSearch. [accessed 17 November 2023].

[379] Bing, no date. Bing safe search settings. [accessed 23 April 2024]

Figure 17.1: Google Search - safe search settings

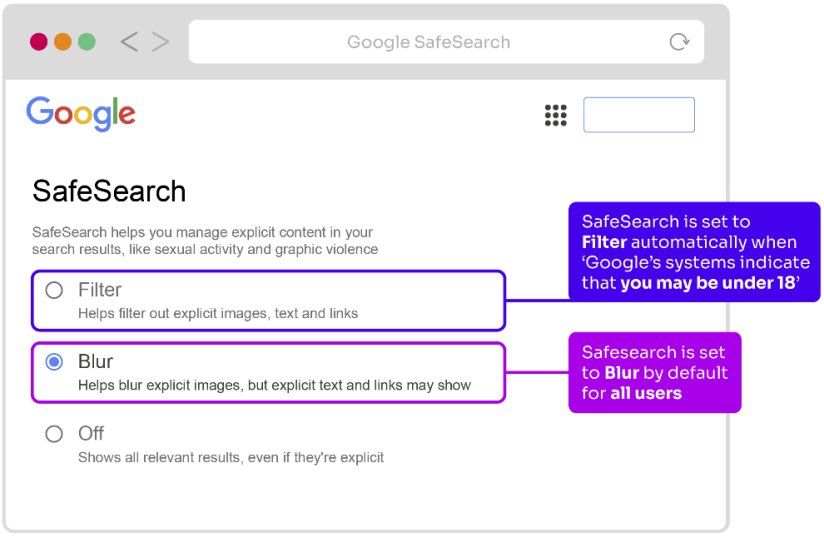
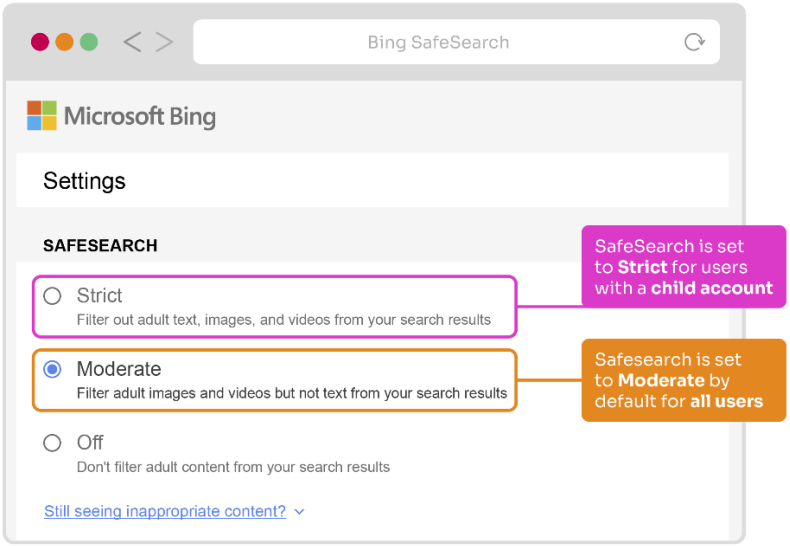


Figure 17.2: Microsoft Bing - safe search settings



17.127 Both Google Search and Microsoft Bing allow users to access and change their safe search settings from any search page within just a few clicks. Google Search provides a safe search toggle that allows users to change the default setting within the search interface and a separate settings page dedicated to safe search. Microsoft Bing provides a separate, direct link to the safe search settings page via a drop-down menu on the search interface. This does not extend, however, to users who have a child account under 13; both services do not allow users with under-13 child accounts to change their safe search settings.[380]

17.128 We note that some smaller general search services (such as DuckDuckGo, Ecosia and Yahoo) also offer safe search settings to filter out adult content, which appear to be set to 'moderate' by default as it relates to adult content.[381] Yahoo's "Moderate" setting filters out

[380]Google, no date. Google For Families Help: Create a Google Account for your child [accessed 12 March 2023]; Bing, no date. Microsoft Support: Parental consent for children's accounts. [accessed 12 March 2023].
[381] DuckDuckGo, no date. DuckDuckGo Settings. [accessed 4 January 2024]; Ecosia, no date. Settings - Ecosia. [accessed 4 January 2024]; Yahoo, no date. Select your setting for Yahoo SafeSearch. [accessed 4 January 2024].

adult content in the form of images or video, it is unclear what formats of content (i.e images/ video/ URLs) are covered in DuckDuckGo and Ecosia's safe search settings. These services generally offer users the ability to change their default "Moderate" safe search setting (including upgrading to "Strict" or switching off) via the "settings" page.

Identifying and including all PPC in safe search

17.129 It is our understanding that services may use automated content detection and manual detection via reporting and complaints to identify content. Content will then be actioned in accordance with the policies in place for each safe search setting.

17.130 We understand that the existing safe search settings of large general search services apply to adult/pornographic content. Google, for example, refers to "explicit" content "like sexual activity" and Bing, DuckDuckGo and Ecosia refer to "adult."[382] It is our understanding that these content descriptions likely broadly correspond to the "pornographic content" category of PPC outlined in the Act.[383]

17.131 We are not aware that services currently apply safe search to other PPC, namely content that encourages, promotes or provides instructions for suicide, self-harm and eating disorders. We therefore recognise that our measure would require in-scope services to expand the current scope of their safe search functionality to include suicide, self-harm and eating disorder content, to ensure that action is taken in respect of all forms of PPC.

17.132 We recognise that extending safe search systems to cover all forms of PPC will require additional efforts on behalf of large general search services. However, given that services have the existing technical framework to implement moderation actions via safe search, we provisionally consider it would be technically feasible and proportionate to recommend they do so to meet their children's safety duties.

17.133 While we understand that the primary tool used by large general search services to operate safe search is automated detection and content classifiers, we do not propose to recommend that services develop any new, or extend existing, automated detection technologies to cover all categories of PPC. We have limited evidence on the technologies currently used by services, or which may be required to extend existing practice to all suicide, self-harm and eating disorder content. Our understanding is it may be difficult to accurately identify suicide, self-harm and eating disorder content through automated means given the complicated nuance surrounding these content areas. For these reasons, and those outlined in 'Other options considered' below, we do not propose to require that services use automated tools to identify PPC for the purposes of safe search. It is open to services to identify content through:

a) review of content identified by user reporting and complaints channels. Given limitations of content detection technologies, we expect this may be the primary means by which search services identify suicide, self-harm and eating disorder content for the purposes of this measure; or
b) automated technologies, including existing technologies that underpin their safe search functions, particularly for the identification of pornographic content.

---

[382] DuckDuckGo, no date. DuckDuckGo Settings. [accessed 4 January 2024]; Ecosia, no date. Settings - Ecosia. [accessed 4 January 2024]; Yahoo, no date, Select your setting for Yahoo SafeSearch. [accessed 4 January 2024].

[383] Pornographic content means content of such a nature that it is reasonable to assume that it was produced solely or principally for the purpose of sexual arousal. See Section 232 (1) of the Act.

17.134 While some existing safe search practices may cover the "violent content" category of PC outlined in the Act (for example, Google Search includes "graphic violence"), at this time, we do not propose to codify this existing practice as part of Measure SM2. This is because we do not think it is proportionate to treat all PC, which includes violent content in the same way that we propose to treat PPC given the lack of evidence of children's experience of violent content on search services, and the costs that services may incur to address this type of PC, in addition to the costs that services will incur to develop their safe search systems to address all PPC, justified below. As we learn more and grow our evidence base, we may consider including this type of content – and other PC harms – in future iterations of our Codes. In the meantime, where services already voluntarily address violent content via their safe search settings, we encourage them to continue doing so.

Safe search settings for users believed to be a child

17.135 As outlined above, we understand that some search services give users the option to share their age. This is primarily through self-declaration on sign-up and, to a more limited extent, through technologies to profile users. We note that Google Search and Microsoft Bing allow users to set up child accounts for children under 13.[384] The child must input their date of birth and their account must be linked to a parent account. Current evidence, including responses to Ofcom's 2023 Protection of Children Call for Evidence, indicates that child accounts are the primary way Google Search is made aware of children on their service.[385] In its response to Ofcom's 2023 Protection of Children Call for Evidence, Google Search notes that it also uses age inference technology to assess user's age. [386]

17.136 Once a child account has been set up, we understand that both Google Search and Bing apply by default Tier 1 (i.e. highest) settings in their respective "safe search" functionalities. Google's Help Center declares that "Filter" (Google's highest safety setting) is the "default setting when Google's systems indicate that you may be under 18."[387] When a child turns 13, or sets up an account at 13, Google Search and Microsoft Bing give users the ability to manage their own account, detach themselves from the linked adult account and change their safe search settings.[388] The default setting for users whose ages are declared to the service to be between 13 and 18 will also have the highest safety setting applied.

17.137 Evidence of current safe search practice suggests that large general search services have existing technical infrastructure that allows for specific safety settings to be applied for certain categories of users that meet certain criteria (i.e. those that have a child account). We therefore provisionally consider that our recommendation that services filter out PPC for users believed to be a child is technically feasible.

Filtering of PPC

17.138 As outlined in Table 17.1 above, 'filtering' involves ensures that content is not returned in search results. We did not consider it proportionate to recommend filtering as an appropriate action in the context of Measure SM1A and Measure SM1B given that the action would be applied for all users, including adults. However, when targeted at users believed to be a child through safe search settings, we provisionally consider that filtering may be a

---

[384] Google, no date. Google For Families Help: Create a Google Account for your child; [accessed 4 January 2024]; Bing, no date. Microsoft Support: Parental consent for children's accounts. [accessed 4 January 2024].
[385] Google response to 2023 Protection of Children Call for Evidence. [accessed 4 January 2024].
[386] Google response to 2023 Protection of Children Call for Evidence. [accessed 4 January 2024].
[387] Google, no date. Filter or blur explicit results with SafeSearch. [accessed 4 January 2024].
[388] Google, no date. FAQs Family link. [accessed 4 January 2024].

proportionate and particularly effective action that will enable services to comply with the duty to minimise the risk of children in any age group encountering PPC.

17.139 Unless an adult opts into the safe search setting where PPC is filtered out, or they are incorrectly determined to be a child (for example a parent using their child's account) we do not envisage them being impacted.

17.140 While there is evidence that default settings are generally effective as users often do not move away from the default setting, we recognise that there is still the option for them to do so where this is allowed by a service. This therefore creates a residual risk that children may turn off the default safety settings. We therefore propose to recommend that when users are believed to be a child, they are not given the ability to turn off their default safe search settings. We acknowledge that this may involve services altering their existing practices, which, as outlined, generally only restrict under-13 users from changing their safe search settings.[389]

Rights assessment

17.141 This proposed measure would require large general search services to go further than we propose under Measure SM1 in respect of PPC for all users believed to be a child in that, rather than simply downranking or blurring identified PPC, they would have to apply a safe search setting for all users believed to be a child which filters out identified PPC from search results. Services would also have to take steps to ensure that this safe search setting cannot be switched off by those users. This measure should be seen as part of the package of search moderation measures that we recommend large general search services adopt, namely Measure SM1A and Measure SM1B for all users. We consider this to be material to our assessment and we have therefore assessed these considerations below.

Freedom of expression

17.142 In addition to the impacts identified in Measure SM1A and Measure SM1B, a potential interference with users' - largely children's - rights to receive information arises in this proposed measure in every case where the service provider has identified PPC and filters that content for users believed to be a child through a safe search setting that cannot be switched off. As a result of the same processes, the freedom of expression rights of interested persons (i.e. website operators) will also be impacted, not only in respect of the PPC identified in URL or image-based results, but also any other non-harmful content hosted at the same URL that will also be filtered out.

17.143 We acknowledge that filtering of URLs or image results that contain PPC constitutes a potentially significant interference with the rights of child users and website operators. As explained above, the act of filtering results in the relevant content no longer appearing in search results. In practice, it means that users believed to be a child will no longer be able to encounter that URL or image-based search results via that service, affecting also any non-harmful content hosted at the same URL. The proposed measure specifies that those users believed to be a child should not be given a means to switch off the setting to access the filtered content.

---

[389] Google, no date. Google For Families Help: Create a Google Account for your child. [accessed 10 December 2023]; Bing, no date. Microsoft Support: Parental consent for children's accounts. [accessed 10 December 2023].

17.144 As with Measure SM1, these impacts have the potential to be significant, particularly if the judgment that the search content is PPC is incorrect. However, the reflections set out in the 'Freedom of Expression' section in Measure SM1, in relation to the incentives of search providers to make correct judgments, and the mitigation provided by the complaints handling processes operated pursuant to section 32(2) of the Act, are relevant. As noted in respect of Measure SM1A and Measure SM1B, the complaints process may also mitigate the impact on the interested persons' right to freedom of expression by giving website operators a mechanism for redress and providing a route to rectify any negative impact by having their content restored to an equivalent position to the one it would have been in had the action not been taken. In addition, the complaints process may also mitigate the impact on any adult users who are wrongly identified as child users in that they would also need to accept complaints from and provide redress to users who have been unable to access content as a result of this measure because of an incorrect assessment of the user's age (for example, by giving them a mechanism to turn off the safe search setting).

17.145 The duty for services to take appropriate action in relation to PPC to minimise the risk of children encountering it is a requirement of the Act. This proposed measure contemplates a more restrictive approach for moderating PPC compared to Measure SM1, requiring that in all cases, services filter identified PPC for users believed to be a child through a safe search setting that cannot be turned off. However, the proposed measure is designed in such a way as to minimise the potential impact on freedom of expression where possible. The measure only involves providers filtering PPC where providers become aware of its presence on the service and does not involve services taking any particular or proactive steps related to content of which they are not yet aware.

17.146 Crucially, the measure only requires the filtering of PPC is applied for users believed to be a child. This is distinct from the approach taken in Measure SM1A and Measure SM1B, in which any action taken by a service is likely to apply to children and adults alike (except where services have reasonable grounds to believe users to be adults). In addition, we chose to recommend the action of filtering as it does not impact the underlying index from which the service presents search results, and, therefore, it will still be possible for all users (other than those which are believed to be a child, and subject to any moderation action taken in line with Measure SM1A and Measure SM1B) to access the content. The filtering of PPC and its benefits of protecting children are therefore narrowly targeted at child users, without impacting the rights of adult users to encounter PPC on the service.

17.147 In addition, we consider this more restrictive approach to PPC to be justified given our current evidence on the risk factors associated with search services (generally, and with regards to certain functionalities), the severe nature of PPC harms to children, and evidence that PPC is particularly prevalent on large general search services.

17.148 To the extent that the actions taken as a result of this measure prevent users believed to be a child from encountering such content on or via search services, we consider that is justified in line with the duties of the Act, as the benefits of the protections on children should outweigh the restrictions on the rights of those children and website operators.

17.149 We therefore consider that the impact of the proposed measure on the freedom of expression rights of the child users it affects and interested persons, above and beyond the requirements of the Act, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this,

and the benefits to children, into consideration, we consider that it is therefore proportionate.

17.150 The proposed measure may also have a narrower impact on services' rights to impart information as, in the narrow case of users believed to be a child, they will now need to take steps to filter identified PPC out of their search results (to the extent that they do not already choose to do so). However, in line with our analysis above, most of this impact arises from the duties placed on services under the Act by the UK Parliament. We, therefore, consider that the impact of the proposed measure on services' rights to freedom of expression is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

Privacy rights

17.151 As explained in Volume 1, Section 2, Article 8 of the ECHR confers the right to respect for individuals' private and family life. For the reasons outlined in our assessment of the impact on the right to privacy in relation to Measure SM1A and Measure SM1B, we do not consider that the process of identifying and filtering PPC for the purposes of this proposed measure would amount to an interference with any user's rights to privacy under Article 8 ECHR.

17.152 We acknowledge that, to implement this measure, services will rely on their existing efforts to assess users that are believed to be children, which may include self-declaration on sign-up, user profiling technologies, or other tools that do not amount to highly effective age assurance. Any tool relied on would likely involve the processing of personal data in relation to users (the nature and extent of which will depend on the precise tool employed). These tools would be needed to form the belief that a user is a child in the first instance, and later to lift the un-changeable setting at such a time as they believe the user to no longer be a child.

17.153 However, the proposed measure does not require that services process any personal data they would not already be processing through whatever means they use to identify users believed to be children. We would expect that any processing of personal data involved in the processes through which services believe users to be a child would comply with relevant data protection legislation. This means that they should apply appropriate safeguards to protect the rights of both children, whose personal data may require special consideration,[390] and adults, where the tools employed involve processing personal data that is not specifically provided by the user (unlike self-declaration).

Impacts on services

17.154 If services do not already have in place the safe search settings and associated actions as proposed, they will need to develop a mechanism which ensures PPC is filtered out for users believed to be a child and take steps to ensure these users cannot turn the setting off.

17.155 As this measure does not require services to proactively identify and classify PPC (for the reasons described above in 'Identifying content harmful to children' in Measure SM1), we have only considered the costs of dealing with the content after it has been identified. Whether services choose to identify PPC through user reporting and complaints channels or

[390] In line with Recital 38 UK GDPR.

through automated technology, they will have to ensure that content flagged as PPC is actioned appropriately according to each safe search setting.

17.156 Services will incur costs to build systems to filter PPC, if they do not already have these systems in place. As assessed in relation to developing blurring and downranking systems in Measure SM1A and Measure SM1B, we believe that most of the cost of to develop a filtering system will come from the quality assurance process. We expect the one-off direct implementation costs for this could be around £7,000 to £18,000, with an annual maintenance cost of £2,000 - £4,500 per annum. There may be additional costs where large services[391] employ significant review, coordination and governance processes in relation to changes of this kind.

17.157 Services may also incur further costs associated with enabling the provision of different search experiences to different users, if they do not already have this capability. This could involve substantial one-off system changes to ensure that the relevant safety settings are applied to users believed to be a child and that the correct actions are applied, but that these actions are not applied to other users. We believe this could require significant resources, as services would need to ensure that the settings apply correctly on all user-interfaces. Services will have to undertake quality assurance, testing whether the settings are applying appropriately based on the type of user. We also expect that for many services to implement this measure effectively, it may be necessary or desirable to allow for user registration if they do not already do so. Although not a specific requirement of the measure, this may entail additional costs.

17.158 Indicatively, we estimate it could take approximately 26 to 39 weeks of software engineering time, with an equivalent amount of non-engineering time, to design, test and implement these safety settings. We expect the cost to develop these settings could be around £58,000 to £170,000. Assuming an annual maintenance cost of 25% of the implementation cost, this would be £14,000 - £43,000 per annum.

17.159 While this measure requires service providers treat users believed to be a child a specific way, we are not requiring service providers to take any additional action such as age assurance to determine which users are children.

17.160 We note that the costs for some service providers may be lower than our estimates where they already have part, or all, of the proposed measure in place to protect children. For example, we are aware that several services, including smaller ones, already have safe search settings in place for some harmful search queries. To the extent that if services' existing safe search settings deviate from the measure, there will be a cost to adapt their settings to comply with the settings we have recommended for different user types, and to extend their settings to cover all PPC.

Which providers we propose should implement this measure

17.161 We propose that this measure apply to all large general search services likely to be accessed by children. We believe that the measure can have important benefits for children's safety online by limiting the extent to which children encounter PPC on these services, which are widely used by children.

[391] See Framework for Codes at Section 13 within this Volume for a definition of a large service.

17.162 Given the risks presented by large general search services and their technical capabilities and resources, we provisionally consider it is proportionate to recommend that they apply the discussed safe search settings which filters out identified PPC for users believed to be a child. We believe such services are likely to have the capacity to implement the measure.

17.163 At this stage, we do not propose to recommend this measure for smaller general search services. The benefits are likely to be materially lower due to the lower reach of smaller services and the fewer children affected. While we are aware that some smaller services already have a tiered approach to safe search for some content, and have filtering and downranking functionalities, we do not know whether the costs are such that it would always be proportionate to recommend this measure for smaller services.

17.164 At this time, we think it is proportionate for smaller general search services to moderate PPC in line with Measure SM1A, which recommends what other services should do upon identifying PPC to meet their Protection of Children duties. They can, of course, choose to implement any action taken in line with Measure SM1A via their existing "safe search" infrastructure if they consider this appropriate.

17.165 We therefore propose that this measure should apply to all large general search services (regardless of risk level) likely to be accessed by children.

Provisional conclusion

17.166 Given the harms this measure seeks to mitigate in respect of PPC, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCS B2 in Annex A8.

Measure SM3: Setting internal content policies

Explanation of the measure

17.167 We recommend that service providers should implement and document clear internal content policies to help ensure consistency, accuracy, and timeliness of moderation decisions.

17.168 We recommend that large general search services and search services that are multi-risk for content harmful to children should set internal content policies that establish rules, standards, and guidelines about what content is, and is not, allowed on the service, and how policies should be operationalised and enforced.

17.169 Services should consider the following when establishing their policies:

a) their most recent children's risk assessment;
b) emerging harms related to content that is harmful to children.

17.170 We believe these factors can help increase the effectiveness of search moderation systems to minimise children's exposure to harmful content.

17.171 This measure builds and expands on the equivalent Illegal Harms measure to apply to PPC, PC and NDC.[392]

[392] See our Illegal Harms Consultation, Volume 4, Section 13.

Effectiveness at addressing risks to children

17.172 We understand that content policies underpin the existing moderation practices on many search services. We consider these to be a necessary step to ensure effective moderation systems are in place for general search services and multi risk services where there is a material risk to children encountering harmful content and to keep users, including children, safer online. Accordingly, we have considered the case for including a measure requiring large general search services and multi risk services to set internal content policies having regard to at least the findings of their risk assessment and any evidence of emerging harms on their service.

17.173 Content policies can serve as an enforcement guide for teams involved in search moderation. As per Section 16, Content moderation for U2U services, content policies set the definitions, examples and exceptions for content allowed and prohibited on a service. As such, content policies can help inform moderation decisions and the design of automated systems trained to identify violating content.

17.174 Internal content policies are typically more detailed versions of external policies; external policies are aimed at users of the service and provide an overview of a service's rules about what content is, and is not, allowed. By setting clear internal content policies, and keeping a written record of these, services can increase the effectiveness, accuracy and consistency of decision making, and reduce the time that content harmful to children remains on the platform. We believe the same is true for search services and search moderation functions.

17.175 Search moderation policies can help to secure more accurate and consistent decision making, particularly for larger or multi-risk services that need to moderate large volumes of diverse content and which may have a large team responsible for content moderation. Though large general search services do not publish internal content policies, evidence indicates existing content policies incorporate PPC or PC to some extent.[393]

17.176 As per Section 16, Content moderation for U2U services, where services consider risk assessments and evidence of emerging harms in setting and updating their internal content policies, there will be considerable benefits to keeping users, including children, safer online. We believe risk assessments and emerging harms can help improve the quality of search moderation policies by pointing to the challenges that moderation functions might face and informing search moderation of PPC, PC and NDC considered harmful to children. By reviewing and updating content policies regularly to reflect these trends, we expect services can improve the quality of their internal content policies and, by extension, improve the performance of their moderation functions. As such, we assess it will be less likely for children to encounter PPC, PC and NDC, and children will have a safer online search experience.

[393] Google Help Centre, no date. Search features policies refers to dangerous, harassing, hateful, medical and sexually explicit content. [accessed 11 December 2023]; Microsoft Bing, no date. How Bing Delivers Search Results refers to adult content. [accessed 5 January 2024].

Rights assessment

Freedom of expression

17.177 This measure builds on the search moderation measures outlined in Measures SM1A, Measure SM1B and Measure SM2; we have not identified any specific additional adverse impacts on the rights to freedom of expression of users, interested persons or services, beyond those already discussed in relation to those measures. This proposed measure is designed in a way that does not tell services how to moderate content that is harmful to children (beyond the actions set out in Measures SM1A, SM1B and SM2), but rather, recommends that there are internal content policies outlining how to moderate it.

17.178 Where services are likely to be dealing with large volumes of search content, the process of considering the scope and application of their content policies in advance would tend to improve internal scrutiny, and improve the consistency and predictability of decisions, in a way which we think would also tend to protect the freedom of expression rights of users and interested persons, and offer more effective protections for children.

17.179 There is some risk that in writing their policies, services which seek to align their publicly available statement with the definitions of PPC, PC and NDC in the Act, may make them of more general application than needed in a way which leads to over moderation (though where they choose to rely on broader definitions, this remains a commercial matter for services). However, we consider that this risk arises equally if we were not to recommend this measure, since content moderators operating without any internal guidance may also over-generalise or be overly cautious.

Privacy

17.180 For the reasons set out above in connection with Measure SM1A, Measure SM1B and Measure SM2, we do not expect this proposed measure would result in any interference with any users' rights to privacy under Article 8 ECHR. Nor do we expect it to involve any additional processing of users' personal data, above and beyond what may already be required for the purposes of Measure SM1A, Measure SM1B and Measure SM2, which we would expect to happen in accordance with data protection legislation.

Impacts on services

17.181 Service providers are expected to incur direct costs if they need to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

17.182 Service providers that do not already have in place internal content policies would incur the full costs of developing such policies. For a smaller search service, developing these policies could take up to three weeks of full-time work and involve legal and regulatory staff, and online safety/harms experts. In some cases, services may use external experts which could increase costs. Engagement and approving new policies may also take up senior management's time, which would add to the upfront costs.

17.183 We estimate that for such services, the one-off direct costs would be in the thousands of pounds. For example, based upon our wage estimate assumptions as set out in Annex 12, if a service required 3 weeks of time across professional occupations (legal/regulatory staff) and 4 hours of senior leadership time, to develop an internal content policy, this would represent a cost of approximately £3,000 to £7,000. However, larger and riskier services may require more complex content policies as the way in which harm can materialise is likely to

be more varied on such services, and the governance requirements needed to implement them are also likely to be more complex. These factors may increase costs given the additional time required to design these more complex policies. These costs could reach the tens of thousands or more. In addition, there may be some small ongoing costs to all services to ensure these policies remain up to date over time (e.g. to take into account emerging harms).

17.184 Some service providers will also be in scope of the related measure proposed in our Illegal Harms consultation.[394] We consider there may be some overlap between the measures, for example, where similar guidelines may apply about how certain aspects of the policies are operationalised and enforced. Any such overlaps and associated cost synergies are likely to be limited given the very different nature of the harms addressed. Likewise, some services will already have policies in place that, at least partly, address this proposed measure. For these services, the proposed measure may mainly involve costs to update existing policies in line with risk assessments and any emerging evidence of PPC, PC and NDC harms.

17.185 We believe that the risk of unnecessary costs are mitigated by the flexibility of the measure. We are not prescribing what should be included in services' internal content policies, but instead propose to set out high-level requirements that give services flexibility to decide how to achieve what is required. This flexibility will allow them to take an approach proportionate to the risks they carry.

Which providers we propose should implement this measure

17.186 We propose that this measure applies to all search services likely to be accessed by children that are multi-risk for content harmful to children regardless of size (which may include vertical services) and all large general search services (regardless of risk level) as these services pose significant risks of harm to children. We consider that the benefits of applying this measure to these services is likely to be material. We are not proposing to apply this measure to smaller general search services which are not multi-risk and vertical search services which are not multi-risk as the benefits will not be as large.

17.187 Large general search services and multi-risk search services of all sizes operate in a more complex risk environment. These services are unlikely to be able to moderate search content effectively without such policies as they need to moderate large volumes of diverse content. As outlined above, the absence of effective search moderation systems and processes significantly increases the risk that children can access content harmful to them on search services.

17.188 The costs of this measure are likely to scale with the number of risks, and so, will scale with the benefits. The cost is likely to be limited for each harm and therefore limited relative to the potential benefits of improving the consistency and quality of content moderation in a complex risk environment. We therefore consider that it would be proportionate to apply the measure to search services that are multi-risk for content harmful to children.

17.189 We also consider that large general search services (of all risk levels) pose significant risks of harm to children and that having internal content moderation policies in place for such services will, therefore, have important benefits for users. We have considered the nature and prevalence of content that is harmful to children can change over time, meaning that even if a large general service is currently low-risk, this could change over a short period in

[394] See our Illegal Harms Consultation, Volume 4, Section 13.

the future. Having an internal content moderation policy in place will help ensure that, if there were to be an increased risk of harm to children on such services, this would be dealt with quickly, reducing the resulting harms, which has a potential to affect many users, including children. The policy may also promote consistency in approach where a service has many moderators, which may be the case on a large service even if low-risk. We also note that large general services are likely to have sufficient resources to develop or adjust these policies in line with the proposed measure. We thus consider that it would be proportionate to apply this measure to large general search services which are not multi-risk.

17.190 As explained in SM1, at this stage we are not proposing to recommend this measure for smaller general search services that are not multi-risk for content harmful to children. Although we propose that such services will need to take appropriate action on content harmful to children (Measure SM1A and Measure SM1B), we consider that it is appropriate to give these services more flexibility given that they are operating in a less complex environment. We therefore consider that the benefits of having a formal, structured framework in an internal content policy would be more limited relative to the costs.

17.191 Our analysis suggests that the benefits of this measure would be materially smaller for vertical search services as these services are inherently less likely to present a significant risk of children encountering PPC, PC and NDC, given they only direct users to content provided by entities with whom they have a direct and ongoing contractual relationship. Therefore, the benefits of having internal content policies are likely to be much lower. We therefore do not propose to extend this measure to large vertical search services just because they are large. However, we believe that this measure would be proportionate if a vertical search service was identified as being multi-risk due to the higher volume of content requiring assessment; we propose that the measure would apply in this case.

17.192 We therefore propose that this measure should apply to search services likely to be accessed by children that are multi-risk for content harmful to children regardless of size (which may include vertical search services) and all large general search services (regardless of risk level).

Other options considered

17.193 When developing our proposed measures, we also considered recommending services extend content policies to apply to human quality raters.[395] We do not have sufficient evidence to suggest that an associated measure would contribute to our aim to effectively minimise the risk of harm to children.

Provisional conclusion

17.194 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm search services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCS B3 in Annex A8.

[395] Google, no date. Search Quality Raters. [accessed 17 November] explains how they evaluate how effectively the service is in delivering content to users.

Measure SM4: Setting performance targets

Explanation of the measure

17.195 We propose that providers of large general search services and search services that are multi-risk for content harmful to children should:

a) set performance targets for their search moderation functions; and
b) effectively track performance in moderation against their set targets.

17.196 We do not prescribe specific performance targets, but suggest, at a minimum, that targets refer to:

a) time that harmful PPC, PC and NDC remains on the service before it is identified and actioned; and
b) accuracy of moderation decisions.

17.197 We believe that performance targets based on time and accuracy will help services act swiftly in response to identified harmful PPC, PC and NDC, whilst balancing this against the desirability of making accurate decisions. Overall, this will help minimise children's exposure to harmful content.

17.198 This measure builds and expands on the equivalent Illegal Harms measure to apply to PPC, PC and NDC.[396]

Effectiveness at addressing risks to children

17.199 Section 16, Content moderation for U2U services, explains that some U2U services currently set performance targets for the operation of their content moderation functions. We are not aware of performance targets used by large general search services regarding the median time to act on content. However, Google Search notes that reporting mechanisms on search are designed to allow users to provide information for Google Search to quickly assess and act where necessary.[397] Microsoft Bing tracks accuracy metrics to monitor moderation effectiveness.[398]

17.200 Where search services are clear about the content moderation outcomes they are trying to achieve, and measure whether they are achieving them, they can better plan how to configure their systems to meet these goals and to optimise the operation of these systems.

17.201 While we do not consider it appropriate to prescribe an exhaustive list of the performance targets that services should include, we consider there are important benefits to services setting both time and accuracy-based targets for their search moderation functions.

17.202 We consider that the children's safety duties imposed on search services imply a need to act swifty in detecting and taking action related to content that is harmful to children, where proportionate to do so. Children will be more effectively protected if decisions are made in a timely way. We therefore think it would be appropriate that time should be included as one of the minimum performance targets required as part of this measure. For example, if there is an unsatisfactory time-lag between identification and action being taken in relation to PPC, PC and NDC, this may have a detrimental effect on children as there is a higher risk that

[396] See our Illegal Harms Consultation, Volume 4, Section 13.
[397] Google response to 2022 Illegal Harms Call for Evidence. [accessed 6 November 2023].
[398] Microsoft Bing, 2023. Bing EU Digital Services Act Transparency Report. [accessed 13 December 2023].

children could encounter the content if it is accessible on the service for a prolonged period. The performance target will ensure services are made aware of their underperformance.

17.203 We are conscious, however, that a focus only on speed and time-based performance targets may result in poor quality decisions.[399] Our measure aims to mitigate this risk by not specifying time targets for services and recommending that services set accuracy targets in addition to time-based performance targets. This will ensure that a focus on speed of decision making is balanced against a focus on accuracy, and that services are made aware if any accuracy rates decline so that they will be in a better position to respond to underperformance. We recognise that services will need to determine the appropriate balance between targets for time and accuracy to help ensure the quality of search moderation practices, and we note that the importance of this balance has been highlighted by several stakeholders;[400] this balance will be subject to the specific risks and needs of each service.

Rights assessment

Freedom of expression

17.204 This measure builds on the search moderation measures outlined in Measure SM1A, Measure SM1B and Measure SM2 above, and we have not identified any specific additional adverse impacts on the rights to freedom of expression of users, interested persons or services, beyond those already discussed in relation to those measures.

17.205 As outlined, the risks to freedom of expression can be increased by the addition of performance targets, particularly targets relating to speed as these can cause moderators to try to take decisions quickly, increasing the risk of error. However, this proposed measure is designed in such a way that requires service providers to balance the need to act swiftly in detecting and taking action in response to content that is harmful to children, with the need to make accurate moderation decisions. In particular, it does not specify a time within which decisions must be made, so the option should not put pressure on moderators to act so fast as to put users' rights to freedom of expression at risk.

17.206 We recognise that there are a range of factors that may affect the likelihood of error (and, therefore, impact on freedom of expression), such as issues with automated technology, turnover of moderation staff, time pressure, and the level of experience of moderators. We consider that the requirement that service providers effectively track their performance against these targets, particularly those relating to accuracy, acts as a safeguard for users' rights to freedom of expression, as against these risks.

Privacy

17.207 For the reasons set out above in connection with Measure SM1A, Measure SM1B and Measure SM2, we do not expect this proposed measure would result in any interference with any user's rights to privacy under Article 8 ECHR. Nor do we expect it to involve any additional processing of users' personal data, above and beyond what may already be

---

[399] Global Partners Digital response to 2022 Illegal Harms Call for Evidence. [accessed 6 November 2023].
[400] Global Partners Digital response to 2022 Illegal Harms Call for Evidence. [accessed 13 December 2023]; Google raised concerns around takedown times in response to Australia's eSafety consultation (2021); Google, 2021, suggested that specifying an exact turnaround time would provide an incentive for companies to over remove content. Australian Government Google submission – Consultation on a Bill for a new Online Safety Act. [accessed 6 November 2023].

required for the purposes of Measure SM1A, Measure SM1B and Measure SM2, which we would expect to happen in accordance with data protection legislation.

Impacts on services

17.208 Service providers are expected to incur direct costs if they need to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

17.209 To implement this measure, a service provider that does not currently have performance metrics and targets in place, would incur both one-off costs to design and set these up, and ongoing costs to track actual performance against established targets. Examples of one-off costs could include creating and implementing processes to track the time between content being reported and when it is assessed and/or action is taken. The flexibility given to services regarding how to implement this measure means that costs are likely to vary widely between services. For example, a service could elect to either build a bespoke ticketing system or license a third-party ticketing system. A simple bespoke system capturing time taken from report to action and estimating accuracy – based solely on the outcome of user appeals – could take approximately a month to design, develop, test and implement. Based on our cost assumptions set out in Annex 12, this is likely to represent a cost of around £8,000 to £16,000.[401] Similarly, off-the-shelf third-party ticketing solutions are available from around £50/month per moderator.

17.210 However, the cost of designing and implementing more complex systems tracking a more extensive set of metrics and carrying out proactive quality assurance of report accuracy would introduce complexity which may significantly impact on the cost. As such, depending on the service design and/or volume of reports, costs could run from the tens to hundreds of thousands of pounds.

17.211 In addition to the initial implementation costs, there would be ongoing costs, which may include data storage costs, and costs to measure and analyse performance against these metrics (e.g., analytics teams). To assess the accuracy of content moderation decisions, services are likely to need to take a sample of those decisions and re-assess them. We have not quantified these costs as they are likely to vary greatly depending on the characteristics of a service. For example, a smaller search service which has a multi risk of two types of content harmful to children may be able to track performance against a single or small number of simple accuracy targets. On the other hand, costs may be significant where services have larger and more diverse kinds of content which pose material risk across many kinds of content harmful to children, potentially requiring more complex and extensive accuracy metrics and greater resource to conduct quality assurance across a large sample of decisions.

17.212 For service providers that are also in scope of the related measure proposed in our Illegal Harms consultation (i.e. services which are large or multi-risk in relation to Illegal Harms), we consider that there may be some overlaps between the two measures due to similarities in the nature of the proposals.[402] The types of metrics and the systems or processes used to track against targets are likely to be similar. Therefore, we expect that the one-off costs associated with the proposed measure will be lower for services that are also in scope of the

[401] Assuming 30 days FTE software engineer time.
[402] See our Illegal Harms Consultation, Volume 4, Section 13.

related Illegal Harms measure. There may be substantial cost overlaps in the ongoing monitoring of performance against these metrics, to the extent that such monitoring is automated, but less so where it is more reliant on human input.

17.213 Some services, particularly larger ones, may already have processes or metrics in place which at least partly address this proposed measure. For these services, the proposed measure may involve any costs of adjusting existing approaches to ensure the recommendations of the proposed measure are met.

Which providers we propose should implement this measure

17.214 We consider that there would be important benefits for large general search services and smaller multi-risk search services (which may include vertical services) from setting performance targets for their search moderation functions and tracking whether they are met. These services operate in a more complex risk environment, and we consider that this measure is particularly important to ensure effective search moderation systems in this context, mitigating the risk of harm to child users. As with Measures SM3 and SM5, these benefits will be greatest for all multi-risk search services (regardless of size) and all large general search services (regardless of risk level) as set out above.

17.215 Although the costs of this measure are significant, we consider that the benefits are likely to be sufficiently important to justify this proposal for all large general search services (regardless of risk level), and multi-risk search services of all sizes, given the fundamental role that effective search moderation plays in protecting users from harm. Large low-risk services may still have significant volumes of cases for moderation, and this measure should help to ensure that, if there were to be an increased risk of harm to children on such services, this would be dealt with quickly and accurately, reducing the resulting harms, which on a large service would have the potential to affect a lot of users, including children. Also, we do not propose to prescribe the details of how services set or achieve the performance targets, leaving scope for services to tailor these targets according to the risks that they identify and the specific operation of their services. This flexibility will help to ensure that services can design performance targets and systems that are proportionate to the risks on the service.

17.216 As explained in relation to Measure SM1, we are not proposing at this stage to recommend this measure for smaller services which are not multi-risk for content harmful to children. We consider that implementing Measures SM1A and SM1B would involve such services having regard to the speed and accuracy of their decisions, but that such services would benefit from greater flexibility in doing so. We consider that the specific approach to performance tracking proposed in this Measure SM4 would not be proportionate for these services as they are likely to face lower volumes of content potentially harmful to children to moderate. Such services may have more limited resources and we consider that the benefit to children's safety may be greater if they focus resources on the core systems and processes for identifying and actioning any harmful content, rather than necessarily investing in additional processes to track performance. We believe that Measure SM1 would provide adequate protection on such services.

17.217 As set out previously, we lack evidence that children encounter PPC, PC and NDC content on vertical services, suggesting that the benefits of this measure would be materially smaller for such services as they are likely to be inherently lower risk. We therefore do not propose to apply this measure to vertical search services at this time just because they are large. We

believe, however, that if a vertical search service is identified as being multi-risk the measure is proportionate, due to the greater risks of harms posed to children and/or the greater volume of content they will need to assess.

17.218 We therefore propose that this measure should apply to search services likely to be accessed by children that are multi-risk for content harmful to children regardless of size (which may include vertical search services) and all large general search services (regardless of risk level).

Provisional conclusion

17.219 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm search services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCS B4 in Annex A8.

Measure SM5: Prioritising content for review

Explanation of the measure

17.220 We recommend providers of large general search services, and search services likely to be accessed by children that are multi-risk for content harmful to children, develop and apply policies on the prioritisation of content for review.

17.221 Services will likely face difficult decisions about what search content to prioritise for review.

17.222 Large general search services facilitate access to large amounts of content and may have to respond to high volumes of reports of potentially harmful content including PPC, PC and NDC. Multi-risk services may also have to respond to high volumes of reports of harmful content given the higher risk posed in comparison to low-risk services and, depending on the service type, might also facilitate access to large amounts of content. We will not prescribe how services should prioritise content review, but suggest that services consider, at a minimum, the following factors:

a) frequency of search requests for the search content;
b) potential severity of the search content, including whether the content is suspected to be PPC or PC or NDC, and the provider's children's risk assessment for the service; and
c) the likelihood that the search content is harmful to children, including whether it has been reported by a trusted flagger, where services use trusted flaggers in their moderation function.

17.223 This measure builds and expands on the equivalent Illegal Harms measure to apply to PPC, PC and NDC.[403]

Effectiveness at addressing risks to children

17.224 Section 16, Content moderation for U2U services, notes that many U2U services use systems and processes to help them prioritise content for review. Services moderating content on a large scale do not typically review content in chronological order, but consider a range of factors, including the virality of the content, its severity, and the context of it becoming

[403] See our Illegal Harms Consultation, Volume 4, Section 13.

known to the platform (for example, whether or not, as a consequence of a user report or other complaint). We anticipate that the same is true for search services.

17.225 We consider that setting a framework for prioritising content review will help search services to identify and prioritise their review of content that presents the greatest risk to children. We believe this will be particularly true for large general search services and multi-risk services, given the amount of content in the indexes they use, or the large volumes of reports they must consider by virtue of their risk level.

17.226 The decisions these services take about what to prioritise can have a material impact on the amount of harm a URL containing harmful content does to children using the service. Effective prioritisation may ensure harmful content is reviewed quickly and, therefore, minimise the risk of children accessing the content. We provisionally consider that the following prioritisation factors are important and relevant to consider:

a) **Search request frequency –** Terms searched more often and by a greater number of users indicate a higher risk of harm to users. Research suggests that systems can be designed to prioritise content according to factors such as the popularity of an item.[404] We are aware that one large search service already considers the frequency with which certain requests are searched by users when prioritising search content for review. Google Search considers factors on the level of harm, including the volume and frequency of search requests.[405]

b) **Potential severity of the search content, including whether it is suspected PPC or PC –** We know that some U2U services already consider the severity of harm when prioritising content for review, and that some harms may be considered to have higher severity than others. We expect that prioritising higher severity search content will help search services minimise harm to users, including children, as such content is likely to pose a more immediate direct harm to the user. Research suggests that systems can be designed to prioritise content according to factors such as the seriousness of the suspected harm, or the likelihood that the item will be confirmed as violative.[406] We believe the same can be applied to PPC, PC and NDC.

c) Likelihood of content being harmful to children - This should include reports by Trusted Flaggers, where available, or content identified as PPC or PC and prioritised based on other identification methods as used by the search service (such as automated systems using natural language technologies). As outlined in Section 16 (Content moderation for U2U services), Trusted Flaggers are any entity for which the provider has established a separate process for the purposes of reporting content which may include content harmful to children, based on the entity's expertise. Though we are not recommending the use of Trusted Flaggers in this iteration of the Code[407] , we know that search services use Trusted Flaggers to identify illegal content. As highlighted in Measure SM1, for example, Google is known to prioritise requests from Trusted Flaggers in relation to illegal content. Section 16 (Content moderation for U2U services) also highlights that signals provided by Trusted Flaggers are particularly crucial in identifying and addressing harmful content in violation of community guidelines on U2U services. We are also

[404] Ofcom, 2023. Content moderation in user-to-user online services. [accessed 6 November 2023].
[405] Google, 2023, Fraud research note to Ofcom.
[406] Ofcom, 2023. Content moderation in user-to-user online services. [accessed 6 November 2023].
[407] Whilst some services currently use trusted flaggers for some illegal content, we do not currently have sufficient evidence on the effectiveness or cost of these programmes to recommend their use more generally for content harmful to children, for full consideration, please see Section 18, User reporting and complaints.

aware that services use other methods to identify harmful material which they may consider to be particularly effective at identifying content with a high likelihood of being PPC or PC; these may include automated content detection technologies. Where services receive flags from Trusted Flaggers relating to relevant categories of PPC and PC, or identify PPC or PC by other, particularly effective content identification processes, we recommend they prioritise such content for review as it could lead to higher quality and more accurate search moderation.

17.227 Overall, we consider that a prioritisation framework which considers the above factors (as well as other factors a service considers relevant) is likely to result in high quality decisions about what search content to prioritise for review. We would expect this to result in a material reduction in harm to children, compared to a counterfactual in which search services simply review content that is potentially harmful to children in a chronological order, thereby delivering significant benefits.

Rights assessment

Freedom of expression and privacy

17.228 This measure builds on the search moderation measures outlined in Measure SM1A, Measure SM1B and Measure SM2, above, and we have not identified any specific additional adverse impacts on the rights to freedom of expression of users, interested persons or services, or on users' rights to privacy, beyond those already discussed in relation to those measures. To the extent that setting and applying a prioritisation policy meant that harm would be a factor in services' decision making and that more users were better protected against harm, it is likely to result in a more proportionate approach to search moderation by the service, and, therefore, would tend to safeguard users' and interested persons' rights to freedom of expression.

Impacts on services

17.229 Service providers are expected to incur direct costs if they would need to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

17.230 Services which do not currently have a prioritisation framework would incur one-off costs to design and set this up (i.e. ensuring that the framework is reflected in systems). We expect these would be largely one-off costs involving a small number of weeks of full-time work and involve legal, regulatory, ICT staff, as well as online safety/harms experts; agreeing on the policy would likely need input from senior management. For example, if designing and setting up a relatively simple prioritisation framework required around three weeks FTE from professional occupations (legal, regulatory, ICT) and one day from senior leadership, this would be equivalent to costs of £4,000 to £7,000 using our salary assumptions as set out in Annex 12. However, for a larger and more complex service with a multitude of different metrics that can indicate virality, severity and suspected type of content, costs could be substantially higher than this, potentially reaching tens of thousands or more, reflecting both more complex design requirements and set-up costs, for example ticketing systems, or systems that automate what content is reviewed next.

17.231 There are also likely to be some smaller ongoing costs to ensure that the prioritisation policy remains reflected in system design, and to review it when appropriate. These costs are

mitigated by the proposed measure not specifying exactly how services should prioritise content, giving services some flexibility in what they do.

17.232 For service providers who are also in scope of the related measure proposed in our Illegal Harms consultation, we consider that there may be some overlaps between the two measures, and the estimated direct costs to these services of implementing this proposed measure would be reduced as a result.[408] For example, metrics related to virality are likely to be similar or the same for both illegal content and content harmful to children. These services will need to consider how they can extend or adapt their existing framework to cover how suspected content harmful to children is prioritised appropriately.

Which providers we propose should implement this measure

17.233 We consider that the benefits of adopting a prioritisation framework for large general search services and multi-risk search services of all sizes (which may include vertical services) are sufficiently important to justify the costs of doing so, given the larger volume and diverse nature of content harmful to children likely to be present on such services. This view is reinforced by the fact that our analysis suggests various services already use prioritisation frameworks of this sort, which is consistent with the costs being proportionate for those services. As the proposed measure does not specify exactly how services should prioritise content, services have flexibility to shape their approach to be proportionate to the risks which are on their service.

17.234 The benefits of recommending this proposed measure to any large general search services that are not multi-risk for content harmful to children will be smaller, as the scope to reduce harm will be more limited. However, similarly to other measures in this section, we still consider that having a prioritisation policy in place for such services will have important benefits for users. Even where a large search service is currently low-risk, this could change over a short period of time (e.g. due to unforeseen changes in their user base or the type of content which is present on their service). Having a prioritisation policy in place will help ensure that services respond efficiently to such circumstances, reducing the resulting harms which, on a large service, would have the potential to affect a lot of users, including children. The policy may also promote consistency in approach where a service has many moderators, which may be the case on a large search service even if low-risk. We also note that large search services are likely to have sufficient resources to develop or adjust these policies in line with the proposed measure. We therefore consider that it would be proportionate to apply this measure to all large services.

17.235 As explained previously in relation to Measure SM1, at this stage we are not proposing to recommend this measure for smaller general search services that are not multi-risk, and vertical search services of all sizes that are not multi-risk. -The benefits of having a prioritisation framework are likely to be materially lower for these services, as they are not likely to need to review as much content of a diverse nature which is potentially harmful to children and are therefore less likely to face difficult consequential prioritisation decisions. The costs to implement this measure (in terms of designing and building a prioritisation framework for such services) could be material, and we do not believe that the potential benefits are large enough to justify these costs to such services. We expect that such

[408] See our Illegal Harms Consultation, Volume 4, Section 13.

services would benefit from greater flexibility in how they organise their content moderation function in the context of measure SM1.

17.236 As set out previously, we believe that vertical search services are inherently less likely to pose significant risks of harm to children, suggesting that the benefits of this measure are likely to be smaller for such services. We therefore do not propose to apply this measure to vertical search services at this time just because they are large. However, in the case that a vertical search service was identified as being multi-risk, the measure would be proportionate, due to the greater risks of harms posed to children and/or the greater volume of content they will need to assess.

17.237 We therefore propose that this measure should apply to search services likely to be accessed by children that are multi-risk for content harmful to children regardless of size (which may include vertical search services) and all large general search services (regardless of risk level).

Provisional conclusion

17.238 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm search services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCS B5 in Annex A8.

Measure SM6: Resourcing search moderation functions

Explanation of the measure

17.239 We recommend that providers of large general search services and search services that are multi-risk for content harmful to children, resource their search moderation functions sufficiently to meet their internal content policies and performance targets. An appropriately resourced search moderation function will help effectively implement search services' search moderation systems and processes.

17.240 We do not prescribe how services should resource their search moderation functions, however, suggest that services take the following factors into consideration:

a) the propensity for significant external events to lead to increased demand for moderation on the service; and,
b) the language needs of services' United Kingdom user base, as identified in its risk assessment.

17.241 This measure builds and expands on the equivalent Illegal Harms measure to apply to PPC, PC and NDC.[409]

[409] See our Illegal Harms Consultation, Volume 4, Section 13.

Effectiveness at addressing risks to children

17.242 We previously stated in relation to the Current Practices associated with Measure SM1 above, our understanding that Google Search and Microsoft Bing use a combination of automated systems and human reviewers in their content moderation functions.[410]

17.243 We are aware that Google Search may use trained experts to manually review and remove content that goes against Google's content policies on a case-by-case basis.[411] Microsoft employs human reviewers to action content based on the service's content policies.[412]

17.244 There is little publicly available evidence about the moderation practices of smaller services. We are aware that Mojeek does not use human moderation in search ranking; rankings are determined by fully automated algorithms based on signals. Mojeek states that content takedown is applied in specific circumstances (i.e. terrorist content, CSAM and spam and malware).[413]

17.245 Responses to our 2023 Protection of Children Call for Evidence stress the importance of adequately resourcing content moderate functions. The Center for Countering Digital Hate (CCDH) suggests that to deliver greater protections for children online, services need to improve their content moderation functions, particularly through substantial resourcing and dedicated human moderators.[414]

17.246 We believe that adequately resourced search moderation teams will better place services to quickly and accurately identify, review, and appropriately action URLs containing potential PPC, PC and NDC according to their internal content policies and performance targets. The measures we propose to recommend regarding prioritisation and performance targets would not protect children unless the service also set out to resource itself sufficiently, and deploy its resources effectively, so as to meet them.

17.247 Our research suggests that moderation resource constraints, and large and fluctuating volumes of potentially violating content can lead to a time-lag between detection and review by moderators.[415] For moderation to be effective, online services may need to quickly scale-up/down operations in response to external events that may cause sudden spikes of illegal content,[416] and build in flexibility. Our research found that services may be able to reduce the turnaround time between content upload and removal by hiring more moderators.[417] Therefore, a search service may be able to reduce how long violative content is returned in search results by adequately resourcing their search moderation teams to respond to fluctuating volumes of content. As such, we consider there would be significant benefits to large general search services and multi-risk search services that consider the possibility of demands for content moderation surging in response to external events, and resourcing their search moderation functions accordingly.

---

[410] Google, no date. Content policies for Google Search. [accessed 20 December 2023]; Microsoft Bing, 2023. Bing EU Digital Services Act Transparency Report. [accessed 20 December 2023].
[411] Google, no date. Content policies for Google Search. [accessed 18 March 2023].
[412] Microsoft, no date. How Bing delivers search results. [accessed 27 February 2024].
[413] Mojeek, no date. Search Content Policy. [accessed 22 March 2024].
[414] CCDH response to 2023 Protection of Children Call for Evidence. [accessed 27 February 2024].
[415] Ofcom, 2023. Content moderation in user-to-user online services. [accessed 19 December 2023].
[416] BSR Response to 2022 Illegal Harms Call for Evidence. [accessed 19 December 2023].
[417] Ofcom, 2023. Content moderation in user-to-user online services. [accessed 19 December 2023].

17.248 We do not think it would be beneficial for us to specify in detail how services should resource their search moderation functions. We do, however, consider that there are factors to which services should have regard when deciding how to resource their search moderation function, which we explain below.

Impact on resourcing of search moderation functions of external events on moderation demands

17.249 We suggest that services take into consideration the propensity for external events leading to a significant increase in demand for search moderation on their service.

17.250 External events may result in spikes of content that is harmful to children online, including harmful, graphic or violent content, which could result in a heightened risk of child users encountering this content if services fail to take proportionate steps to plan for this. Business for Social Responsibility has stressed to Ofcom the importance of all online services being able to quickly scale up/down operations in response to significant events that may cause sudden spikes of illegal content,[85] and Ofcom considers that the same can be said for content that is not illegal but is harmful to children. Services which have contingency plans in place to ensure that content across the system is dealt with expeditiously are more likely to protect children effectively.

17.251 As such, we assess that search services can minimise the risk to child users by similarly being able to scale up operations in response to events that may result in an increase of content that is harmful to children on their services.

17.252 Information obtained from platform risk assessments, tracking signals of emerging harm and other relevant sources of information, could be used to understand where and when such occurrences might happen.

Language proficiency

17.253 It is also important that moderation systems can handle different languages and understand specific cultural contexts to accurately and quickly identify and action content. We know users in the UK use search services in multiple languages.[418] Section 16, Content moderation for U2U services, notes that deploying appropriate language resource and expertise can enable services to identify, review and moderate search content that is suspected to contain content that is harmful to children, and can impact the speed with which content related to specific harms and from specific countries is reviewed on U2U services.[419] We suggest this finding is also true for search services.

17.254 We do not propose to prescribe the specific language expertise or resource required. The language expertise required to deal with the risk of harm will differ per service and will depend on factors including user base and the search service type. We suggest, however, that where search services factor language proficiency into the resourcing of their moderation functions, this is likely to deliver important benefits.

---

[418] Vox, 2015. In which language do you Google? Tracking 135 languages in 9 cities since 2004. [accessed 10 December].

[419] Ofcom, 2023. Content moderation in user-to-user online services. [accessed 19 December 2023].

Rights assessment

Freedom of expression and privacy

17.255 This measure builds on the search moderation measures outlined in Measure SM1A, Measure SM1B and Measure SM2, above, and we have not identified any specific additional adverse impacts from this proposed measure regarding resourcing the search moderation function appropriately on the rights to freedom of expression of users, interested persons or services, or on users' rights to privacy, beyond those already discussed in relation to those measures.

Impacts on services

17.256 Service providers are expected to incur direct costs if they would need to make changes to apply the proposed measure. We have not identified any specific indirect costs related to this measure.

17.257 The total ongoing cost of resourcing services' content moderation functions in line with this measure is likely to be substantial, particularly for larger and riskier services with large volumes of relevant search content to moderate. Whilst many services would in any case have some level of resource allocated to search moderation, a higher level of resources may be required to fully give effect to the policies and targets set out in Measure SM3 and Measure SM4.

17.258 We expect that the level of resource required to implement the proposed measure will vary by size of service and depend upon the policies they develop, and the nature and volume of harmful content present on their service. In general, we would expect costs to be higher for larger general search services, as larger services will tend to have a higher volume of content to review and, therefore, require more resource. However, it is possible that some smaller services, and large vertical services may still face high costs if they are high-risk and therefore have a large quantity of content requiring review. It is for services to consider the level and types of resource required to meet this measure, and to what extent this may entail additional resource and cost.

17.259 For service providers who are also in scope of the related measure proposed in our Illegal Harms consultation, we consider that there may be some limited overlaps between the two measures.[420] For services which are already resourcing their content moderation systems to give effect to internal content policies and performance targets relating to illegal harms, these costs may be somewhat reduced in cases where there are synergies between the two types of content moderation, for example, where a piece of content is both illegal content and content harmful to children. It is also possible that the same resources could be used to review both suspected illegal content and content harmful to children, which could help to manage costs in some cases (e.g. when there is a peak in prevalence of one particular kind of content).

17.260 In all cases, the magnitude of costs is likely to be further influenced by the type of review processes used:

a) For example, automating search moderation processes require both one-off infrastructure investment, and different ICT professionals' time. Larger services may be able to develop these in house, but the costs of doing so can be high. Due to this,

[420] See our Illegal Harms Consultation, Volume 4, Section 13.

smaller services may outsource development to a third party, or use off-the-shelf third-party solutions.[421] In addition, system updates and licensing costs can be expensive and add to ongoing costs.

b) Human moderation resourcing costs will primarily depend on how many moderators are needed. In addition, for search moderation resources to be effective in meeting policies and targets, human moderators may require specific training (see Measure SM7 below).

17.261 There are likely to be trade-offs to services between investing in automated moderation and human moderation, to an extent.

Which providers we propose should implement this measure

17.122 This proposed measure is linked to, and would be effective for, those services which have search moderation policies and performance targets in accordance with Measures SM3 and SM4. This measure is important for those search moderation measures to have the intended effect.

17.123 Our analysis suggests that this measure could impose significant costs on services. However, we consider that where search moderation functions are well-resourced this will deliver very significant and important benefits. We would expect this to ensure that services give effect to Measures SM3 and SM4, which together will result in a material reduction of harm to children compared to a counterfactual scenario where the service operates on lower level of resources that may be insufficient to fully implement their internal moderation policies and achieve targets.

17.124 The costs of this measure are likely to scale with the size and risk level of a service, as larger services and those with a higher risk of hosting content harmful to children will have a larger volume of content to review and therefore higher costs. However, the benefits of such content being identified, and action taken regarding it, will also be higher and we therefore expect that the costs will scale with the benefits.

17.125 We propose to apply this measure to search services of all sizes that are multi-risk for content harmful to children (which may include vertical search services) and all large general search services. As we are proposing that Measure SM3 and Measure SM4 would also apply to all multi-risk search services and all large general search services, it follows that it is proportionate that this measure apply to these services too, to ensure that these proposed measures are effective and able to reduce harm.

17.126 This measure relates to resourcing well aspects of search moderation functions defined in Measures SM3-SM5, which do not at this stage apply to smaller general search services and vertical search services which are not multi-risk. For this reason, we also do not recommend this measure SM6 for smaller general search services and vertical search services which are not multi-risk. However, we note that these services should, in any case, ensure that they have adequate resources to enable them to give effect Measure SM1, even if we give more flexibility as to how they achieve that.

17.127 We therefore propose that this measure should apply to search services likely to be accessed by children that are multi-risk for content harmful to children regardless of size

[421] Pre-built solution offered by a third-party vendor.

(which may include vertical search services) and all large general search services (regardless of risk level).

Provisional conclusion

17.128 Given the harms this measure seeks to mitigate in respect of PPC, PC and NDC, as well as the risks of cumulative harm search services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCS B6 in Annex A8.

Measure SM7: Appropriate training and materials for search moderation

Explanation of the measure

17.129 We recommend that people working in search moderation should receive training and materials that enable them to effectively action content policies, moderate content and improve outcomes for users.

17.130 We do not currently consider that the measure would apply to those voluntarily working in search moderation. We are unaware of any search services that employ volunteer moderators and, therefore, do not envisage this will currently impact any service.

17.131 We recommend that in providing training materials, services should have regard at least to the following matters:

a) the most recent children's risk assessment of the service and information pertaining to the tracking of signals of emerging content that is harmful to children; and
b) where the provider identifies a gap in a moderator's understanding of a specific kind of content that is harmful to children, it gives training and materials to remedy this.

17.132 We consider that training materials may include parts of a service's internal content policies, enforcement guidelines, and examples and individuals of the tools or interface moderation staff will use to carry out their job. We will not recommend how training materials should be developed or updated, or how trainings should be delivered. We assess this will differ from service to service, depending on the service type and user base, and as such, we believe that services are best placed to determine what training and materials are appropriate for their respective services and teams.

17.133 This measure builds and expands on the equivalent Illegal Harms measure to apply to PPC, PC and NDC.[422]

Effectiveness at addressing risks to children

17.134 There is limited evidence on how search services train staff (including contractors, etc.) involved in content moderation. We know that some larger services train their moderators and other relevant members of staff to identify and action violative content, as well as providing supporting materials to help them do so. In particular, we know that Microsoft Bing ensures human reviewers receive extensive training on their policies.[423]

[422] See our Illegal Harms Consultation, Volume 4, Section 13.
[423] Microsoft, 2023. Bing EU Digital Services Act Report. [accessed 20 December 2023].

17.135 Section 16, Content moderation for U2U services, cites evidence from civil society and academics stressing the importance of training U2U moderators, concluding that training staff involved in moderation, and providing them with relevant materials, is beneficial. Staff trained on how to identify and action content harmful to children are more likely to be equipped with the knowledge and skills to do so, compared to untrained staff. This is particularly true where staff are trained regularly to ensure they have up-to-date knowledge of content moderation policies, as well as on the systems they are using to carry out their job.

17.136 We believe that providing appropriate and updated training and materials for staff involved in search moderation will also enable more accurate and informed search moderation decisions. We believe that adequately trained staff involved in moderation can likely make more informed, quicker, and accurate moderation decisions. Overall, this is beneficial for identifying and minimising the risk of children encountering URLs that contain PPC, PC and NDC, especially when compared to not training staff. Staff involved in moderation who are trained regularly will have up-to-date knowledge of content moderation policies, as well as the systems they are using to carry out their job.

17.137 Section 16, Content moderation for U2U services states that U2U content moderation functions include content moderators, including outsourced moderators, and other staff. We expect that the people working in search moderation would also mostly include content moderators employed or contracted by providers. It may also involve others involved in the wider content moderation ecosystem, such as: Trust and Safety staff; quality assurance and compliance staff; subject matter experts; lawyers and other legal staff; risk management staff; operations staff; engineers; and developers.

17.138 While we are aware that some U2U services rely on volunteer or community moderators (see Section 16, Content moderation for U2U services), we are not aware that search services currently employ volunteers to support their moderation functions, and as such, have not proposed that the measure should apply to any volunteers to the extent that they might be used for moderation purposes.

17.139 We do not propose to recommend how training materials should be developed or updated, or how training should be delivered and how frequently. We believe that services are best placed to determine what training and materials to provide and the occurrence of training to respond to the specific needs and risks of their service and staff functions.

17.140 However, services which do not have regard to certain factors are unlikely to protect children properly. We therefore provisionally consider that the matters outlined below are likely to be relevant to prepare and deliver search moderation training and materials:

a) **Risk Assessment**: A service's children's risk assessment is likely to be a key source of guidance for those supporting a services' moderation functions; they can reflect current trends related to search content that is potentially harmful to children that may exist on their service and how it manifests. As noted in Measure SM3 above, risk assessments can form the basis of a service's internal content policies; we assess that where services do consider risk assessments in their internal content policies, they can improve the quality of their search moderation efforts. As moderators should be focused on enforcing the internal content policies, we provisionally consider it crucial that training and materials should also be informed by the most recent children's risk assessment.
b) **Address gaps in moderation staff's understanding**: There may be instances where staff do not have the appropriate or sufficient understanding of specific harms to enable

them to effectively minimise the risk of children encountering content that is harmful to children. Harms-specific training and materials may be helpful to identify and action search content that is harmful to children due to the complex, nuanced nature of PPC, PC and NDC. Specific training should be provided to those involved in content moderation of such content. If training and materials are given to moderators where a service has identified a gap in moderators' understanding of a specific harm, and where they deem there to be a specific risk, this should improve outcomes for children.

Rights assessment

Freedom of expression

17.141 This measure builds on the search moderation measures outlined in Measure SM1 and Measure SM2 above, and we have not identified any specific additional adverse impacts from this proposed measure regarding providing appropriate training to search moderation staff on the rights to freedom of expression of users, interested persons or services.

17.142 Our assessment of freedom of expression and privacy rights impacts associated with having a search moderation function is set out above in relation to Measure SM1 and applies equally in relation to this measure. As several respondents to the 2022 Illegal Harms Call for Evidence noted, training enables those involved in content moderation to make better decisions.[424] [425] Training also enables staff involved in moderation to have a better understanding of borderline content (i.e. content where it can be difficult to determine whether it is legal or illegal). All things being equal, better training should safeguard right to freedom of expression of users and interested persons.

Privacy

17.143 Our assessment on the impact of the right to privacy associated with Measure SM1A, Measure SM1B and Measure SM2, namely that there is no interference, applies equally in respect of this measure. In addition, we note, that to the extent that services choose to use specific items of content in their training and materials, they would need to comply with privacy and data protection laws as outlined relation to Measure SM1A and Measure SM1B. However, our proposed measure does not require them to do so.

Impacts on services

17.144 Service providers are expected to incur direct costs if they need to make changes to apply the proposed measure. We have not identified any specific indirect costs relating to this measure.

17.145 We note that the costs for some service providers may be lower than our estimates below, as some service providers may already have part, or all, of the proposed measure in place. As set out above, we know that many services already provide some form of training to their content moderators.

---

[424] Several respondents to our 2022 Illegal Harms Call for Evidence stressed the importance of training. Wikimedia response to 2022 Illegal Harms Call for evidence.

[425] In addition, 5Rights response to 2023 Protection of Children Call for Evidence; Refuge response to 2023 Protection of Children Call for Evidence; Glitch response to 2023 Protection of Children Call for Evidence; Global partners Digital response to 2023 Protection of Children Call for Evidence; Samaritans response to 2023 Protection of Children Call for Evidence.

17.146 For a service provider to implement the measure, it would incur two main types of cost. Firstly, the costs to develop the training material, including both upfront costs and ongoing costs to keep this updated. The second, are costs to deliver training to moderators. Services which are not in scope of the related measure proposed in our Illegal Harms consultation, would incur the full costs of developing the training material, an estimate of which is included within the cost estimates below.

17.147 The costs associated with delivering the training to content moderators will be impacted by the chosen format of training (e.g. delivered by a human trainer each time or via a video/interactive interface, or on-the-job training), and will also depend upon the number of staff to be trained and the training duration. We assume that content moderators will not be available to perform their usual role during the training process but will be paid.

17.148 We assess these costs to be comparable to those for U2U services content moderation, (see Section 16, Content moderation for U2U services). In summary, we estimate that the costs to provide training for one new content moderator could be between £2,900 and £18,000, and for a new software engineer between £4,700 and £28,000.[426] As the number of moderators that need training is likely to depend on the volume of content that needs to be assessed, the costs of this measure are likely to scale with the benefits. There will also be some ongoing costs for refresher training and training in new harms on the services. We expect the annual costs of these to be lower.

17.149 For service providers who are also in scope of the related measure proposed in our Illegal Harms consultation, we consider that there may be some limited overlaps between the two measures.

17.150 While the kinds of harms and associated content are not the same, services may need to make changes to training content and duration to comply with the children's safety duties so that training is adequate for the OS regime. We therefore expect that platforms who already have training in place to cover these harms, will have slightly lower costs as a result of this measure than those who have no training in place for content moderators at all.

17.151 All other things being equal, smaller services will have less content to review and smaller search moderator teams and, therefore, will incur lower training costs. While costs for services will scale with the risk of harm, this will come with a commensurate benefit. In general terms, we would expect costs to vary with the potential benefits, in the sense that services with higher risk of hosting content harmful to children are likely to need more search moderators and require their moderators to be trained on more harms, therefore, resulting in higher training costs. However, these services are likely to have more search content harmful to children and thus higher benefits from having well trained moderators who can take effective action regarding this content.

17.152 These costs are also mitigated by the fact that this measure does not specify exactly how services should provide training to content moderators, giving services some flexibility in what they do. Services can decide the most appropriate and proportionate approach to

[426] This is based on our assumptions on wage rates set out in Annex 12. We also assume that the wage cost of the people being trained represents only half of the total costs of the training. This is consistent with the Department for Education saying that the wage cost of staff being trained accounted for about half of all training expenditure in 2019, although this varies by the size of the firm and the sector. We assume this excludes the [22%] uplift that we have assumed elsewhere for non-wage labour costs, so we have not also increased these wages by 22%. Source: Department for Education (DfE), Employer Skills Survey 2019: Training and Workforce Development 2020, pp38 and 40. [accessed 5 February 2024].

training search moderators for their own contexts. This flexibility allows an approach that is cost-effective and proportionate for each service.

Which providers we propose should implement this measure

17.153 This proposed measure is linked to, and would be effective for, those services which are recommended to have search moderation policies in accordance with Measure SM3. We consider that for the internal policies measure to be effective, it is necessary for the moderators to be able to identify and action search content according to the internal policies. Though this will be an additional cost, it is important that services are able to identify content harmful to children. It follows that this measure should only be considered for those services which have internal search moderation policies as set out in Measure SM3.

17.154 We consider the benefits of this measure are likely to be high. This is because search moderator training is important in effectively implementing a service's search moderation policies to reduce harm and comply with its online safety duties. Well-trained and prepared search moderators are more likely to be able to identify content harmful to children and, under the service's content standards, apply the correct action to take to it, reducing the harms that result. As the number of search moderators that need training is likely to depend on the size of the service and the volume of content that needs to be assessed, the costs of this measure are likely to scale with the benefits. As such, this measure is likely to be proportionate for services which identify significant risks of harm to users.

17.155 We consider this to be the case for both multi-risk search services of all sizes (which may include vertical services) and large general search services (irrespective of risk). Training costs are likely to depend primarily on the number of people that need to be trained. All other things being equal, smaller services are likely to have smaller volumes of content, and fewer content moderators as a result. This means that the costs for smaller services will be correspondingly lower than for large services.

17.156 As per Measures SM3 to SM6, we also consider the Measure SM7 is proportionate for large general search services that are not multi-risk for content harmful to children. Large services are typically more complex and may have a large volume of content moderation cases even if there is low risk. We consider there is a material potential benefit from appropriate training under this measure, even for such services, mitigating the risk of content moderation failures which could affect a large number of users, including children.

17.157 Smaller general search services which are not multi-risk for content harmful to children, and vertical search which are not multi-risk for content harmful to children are likely to moderate lower volumes of search content that may be harmful to children and the likely scale of harm is much smaller. Therefore, at this stage we do not consider that it is proportionate to recommend this measure for these services. It is expected that these services would need to consider appropriate steps to equip moderation staff to be able to implement Measure SM1, but for such services we are not recommending formal training with the specific elements set out in this measure, therefore providing more flexibility to such services.

17.158 We therefore propose that this measure should apply to search services likely to be accessed by children that are multi-risk for content harmful to children regardless of size (which may include vertical search services) and all large general search services (regardless of risk level).

Provisional conclusion

17.159 Given the harms this measure seeks to mitigate in respect of PPC and PC, as well as the risks of cumulative harm search services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCS B7 in Annex A8.

Other issues to note

17.160 As discussed in Section 16, Content moderation for U2U services, we recognise the significant impact that human moderation of content can have on the wellbeing of an individual and the importance of providing appropriate supervision and support in this area. We note, however, that the responsibility towards employed moderators is within the employers' remit and, therefore, would only be relevant to our remit if it impacted on user safety. We welcome evidence from stakeholders on this, to which we would have regard in planning our work on future iterations of our Codes.

Other options considered

17.161 In addition to our proposals, we considered whether to recommend proactive technology measures relating to content moderation. We understand that services may use proactive technology, including keyword detection technology, to identify harmful content as part of their content moderation systems, including but not limited to the "safe search" features of many general search services. We considered whether to recommend measures in this space, including that services implement keyword detection and maintain keyword lists, including codewords, to protect children from harmful content. While we are aware this technology exists and some services may implement it to complement their content detection and moderation systems, we require additional evidence of how accurately and effectively it could be used to identify PPC, PC and NDC in search content before proceeding with a specific proactive technology recommendation. We may consider potential measures in this space in future iterations of the Children's Safety Codes, particularly to complement the safe search settings recommended in Measure SM2. We encourage stakeholders to share information about their current use of this technology, associated costs and risks and effectiveness at minimising risk of harm to children.

18. User reporting and complaints

User complaints are an important mechanism for making service providers aware when something goes wrong on their services, such as harmful content being present, or content being mistakenly removed or restricted. As such, complaints play a crucial role in both keeping children safer online and protecting users' rights.

Our evidence shows that while many services have content reporting tools or complaints functions for users, these are not always accessible, easy to use and transparent. This can discourage people from complaining, particularly children. The Online Safety Act 2023 ('the Act') places duties on all service providers regarding the design and operation of complaints processes. In our Consultation: Protecting people from illegal harms online ('our Illegal Harms Consultation'), we proposed a number of measures to help providers meet those duties (see Section 16, Reporting and complaints). In this consultation, we are proposing further measures to help providers of services likely to be accessed by children to meet their duties relating to children's reporting and complaints. We think these measures will also help providers meet their safety duties to protect children, by reducing barriers to complaining for users (including children) and thereby increasing the volume of high-quality complaints they receive. This will help providers to identify harmful content and take steps to protect children from it.

We are also proposing additional measures which our evidence has shown could help protect children and other users from both content harmful to children and illegal content. We are proposing to include these measures in the draft Illegal Content Codes and the draft Children's Safety Codes, as explained in greater detail below. We have assessed the potential impacts of our proposals, including costs and rights impacts, and deem them proportionate for the services indicated.

Our proposals

#	Proposed measure	Who should implement this[427]
UR1	**Have complaints processes** which enable people to make relevant complaints for services likely to be accessed by children	All Search and U2U services
UR2	Have **easy to access and use, and transparent complaints systems**	
UR3	**Acknowledge receipt of complaints** with indicative timeframe and information on resolution	
UR4	**U2U services take appropriate action in response** to each complaint[428]	All U2U services
UR5	**Search services take appropriate action in response** to each complaint[429]	All Search services

[427] These proposed measures relate to providers of services likely to be accessed by children.
[428] For some complaints we make different recommendations for some services, based on their size and risk level. See the measure below for more detail.
[429] For some complaints we make different recommendations for services based on their size and risk level. See the measure below for more detail.

Consultation questions

43. Do you agree with the proposed user reporting measures to be included in the draft Children's Safety Codes? Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence. 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.
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? Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence. 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.
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)? Please provide any arguments and supporting evidence.

The importance of complaints processes for protecting children

Definition box 1: What is the difference between user reports, appeals and complaints?

18.1 **User reports** and **appeals** are types of **complaint**. Throughout this section, we use **complaints** to refer to all types of complaints, including user reports and appeals.

18.2 **User reports** are a specific type of complaint about content, submitted through a reporting tool.

18.3 **Appeals** are complaints by users who believe that their content has been wrongfully taken down or restricted, their account wrongfully suspended or banned, or (for website owners) their content no longer appears in search results.

18.4 Enabling users to make complaints can help ensure services are safer for children, accountable and respect users' rights. The types of complaints mentioned in the Act can be split into two main categories:

a) **Content-related complaints** are important for making providers aware of content harmful to children present on, or available via, their services, which their other content moderation systems have not already identified. They help providers to take steps to protect children from this content or prevent them from encountering it. This reduces the risk of other children being harmed in future. In some cases, providers may also choose to use information from complaints to help refine any automated systems they use to detect harmful content. This enables those systems to identify content harmful to children more accurately. Monitoring content-related complaints can also help providers identify emerging risks developing on their services. This is one reason why it is so important that low-risk services also have effective complaints processes.

b) **Complaints about non-content concerns** play an important role in protecting users' rights and ensuring they are treated fairly. This might include, for example, appeals about content moderation decisions, or complaints by users who cannot access content because their age has been incorrectly assessed.

Definition box 2: Who should be able to complain?

On both U2U and search services, all **users and affected persons** should be able to make complaints about content/search content harmful to children or about the provider not complying with their duties. That includes adults and children using the service (whether or not they are registered) and people with a characteristic targeted by content on the service, as well as parents/carers of children and vulnerable adults who use the service or are the subject of content on the service. We consider teachers of children using the service or the subject of content on the service would also be affected persons.

On U2U and search services, **users**, both children and adults, should be able to complain if they are unable to access content because of an incorrect assessment of their age.

On U2U services, **users** should be able to submit appeals about their content being taken down or restricted for being considered content harmful to children, or their account being suspended or banned for generating content considered harmful to children.

On search services, **website owners** (called 'interested persons' in the Act) should be able to appeal if their content no longer appears or is given a lower priority in search results, for being considered content harmful to children.[430]

Reasons why complaints processes are underused

18.5 While many services have reporting tools or complaints processes, evidence suggests that people, and particularly children, face barriers to using them. Our detailed evidence on this in Section 7.11, Governance, systems and processes, highlights the following main themes:

18.6 Reporting tools can be **difficult to find or not clearly identifiable,** discouraging people from complaining. For example, in 2023, our VSP tracker survey of adults and children aged 13 and over found that 14% of respondents who had been exposed to harmful content tried to use a reporting mechanism but could not find it.[431]

18.7 Complaints systems can be **too burdensome or involve too many steps** and may be **complicated and difficult to understand**. In our 2024 research into children's experiences of violent online content, children who had reported previously frequently said they were discouraged from reporting again because the process was time consuming and complicated. Some said they thought reporting forms were designed for adults rather than children.[432]

18.8 Some providers are **not sufficiently transparent** about how their complaints processes work, leading to lack of trust that they take action in response to complaints. Specifically, our evidence indicates some children are discouraged from reporting because they do not have

[430] The right to have an appeal considered arises for 'interested persons', defined in section 227 (7) of the Act. For readability, in this section we have used the term 'website owners' as shorthand to refer to interested persons.
[431] Ofcom 2023. VSP Tracker Wave 4 data tables, table Q11. Illustrative graph shown on VSP Tracker Wave 4 Chart Pack.
[432] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.

a clear understanding of how to make a complaint.[433] It also suggests they are concerned about the person they complain about finding out.[434] Evidence also indicates that children do not believe service providers do anything meaningful in response to complaints.[435] Not being informed of the outcome of their complaints further reduces children's trust in complaints processes.[436]

18.9 While we recognise that different groups of children may have different needs and concerns when making complaints, our evidence suggests that these barriers discourage many children from complaining. This may lead to content harmful to children being available on services for longer periods of time or not coming to providers' attention at all, if their content moderation systems do not identify the content proactively. It can also leave other problems going unchecked, such as incorrect restriction of content or inaccurate age assessments, which can pose risks to the rights of users by restricting their freedom of expression, including their right to receive and impart information or their freedom of association.

Interaction with Illegal Harms

18.10 In our Illegal Harms Consultation, we proposed the following measures regarding reporting and complaints be included in our draft Illegal Content Codes:

- **Measure 5A:** Providers of all U2U and search services should have complaints processes which enable UK users, affected persons and (where relevant) interested persons to make complaints relevant for all services (as set out in Sections 21(4) and 32(4) of the Act).
- **Measure 5B:** Providers of all U2U and search services should have easy to find, easy to access and easy to use complaints processes.
- **Measure 5C:** Providers of all U2U and search services should acknowledge relevant complaints for all services with indicative timeframes for deciding the complaint.
- **Measures 5D-H (U2U):** Providers of all U2U services should take appropriate action in response to complaints.
- **Measure 5D-H (search):** Providers of all search services should take appropriate action in response to complaints.
- **Measure 5I:** Providers of all large services with a medium or high risk of fraud should establish and maintain a dedicated report channel for fraud, for trusted flaggers.

18.11 See Section 16, Reporting and complaints, of our Illegal Harms Consultation for a detailed discussion of the evidence and impacts of those measures.

---

[433] Ofcom, 2024. Experiences of children encountering online content promoting eating disorders, self-harm and suicide. Ofcom's 2023 Children's Media Literacy Survey also found that 36% of 12-17s who go online said that they knew how to use a reporting or flagging function' (See Children's Online Knowledge and Understanding Data tables).

[434] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK; Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.

[435] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK; Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.

[436] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK; Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.

18.12 We understand that many providers operate a single complaints process for various types of complaints. Children, parents, and other complainants may use this to complain about suspected illegal content, content harmful to children as defined in the Act or other issues falling outside the scope of the safety duties for services likely to be accessed by children. Until the provider reviews the complaint, it will not know what it is about. We have taken all this into account when assessing the impact of recommending measures for inclusion in our draft Children's Safety Codes.

18.13 We provisionally consider that proposed measures 5A-H (for both U2U and search services) in the draft Illegal Content Codes are also proportionate for providers of services likely to be accessed by children in relation to the additional specific types of complaint the Act requires them to handle. We set out below our detailed assessments of the evidence and impacts of these measures as they relate to duties for services likely to be accessed by children. We explain in the 'Further measures considered' section below why we are not at this time recommending dedicated reporting channels or trusted flaggers for content harmful to children.

18.14 Measures UR1, UR4 and UR5 are in substance unchanged from our provisional recommendations in the draft Illegal Content Codes. Where relevant, we have relied upon updated evidence and considered the specific rationale as to why we propose these to be relevant for the additional types of complaints that providers of services likely to be accessed by children are required to handle. In places we have also updated the language used, to improve the clarity of our recommendations.

18.15 We are also proposing that measures 5B and 5C in the draft Illegal Content Codes be included in the draft Children's Safety Codes. However, we are proposing to add additional elements for these measures to protect children from harm, in light of new evidence regarding the barriers children face to reporting. This evidence does not relate to a specific harm or type of content but rather to children's likelihood to use reporting and complaints processes to make any kind of complaint. Due to this, we provisionally think these additional recommendations could help protect children and other users from both content harmful to children as defined in the Act and illegal content.

18.16 We are therefore consulting on including these additional recommendations in both the draft Children's Safety Codes and the draft Illegal Content Codes. We set out details of these measures below. Subject to consultation responses, we will aim to include the additional illegal harms measures in our Illegal Harms Statement.

Our proposals to protect children

18.17 The Act requires providers of services likely to be accessed by children to operate systems and processes that allow users – including children, parents, and carers – to easily report content harmful to children and make other types of complaints. Service providers are also required to take appropriate action in response to complaints and ensure they are transparent about complaints processes and handling.

18.18 The following kinds of complaints are in scope for providers of services likely to be accessed by children:

- complaints about content harmful to children,

- complaints about a user's content or account being removed or restricted (U2U) or a website owner's content no longer appearing or being given a lower priority in search results (search) (we refer to these as appeals),
- complaints about a user's access to content being restricted based on incorrect assessment of their age, and
- complaints about service providers not complying with their duties to protect children.[437]

18.19 We propose five measures in this section. We discuss our detailed rationale for these measures and which services we propose they should apply to later in the section. Our proposals can be summarised as follows:

a) **Measure UR1:** Providers of all U2U and search services likely to be accessed by children should have complaints processes which enable people – including children (and other users), parents/carers and website owners – to make relevant complaints for services likely to be accessed by children.
b) **Measure UR2:** Providers of all U2U and search services likely to be accessed by children should have easy to access, easy to use and transparent complaints processes.
c) **Measure UR3:** Providers of all U2U and search services likely to be accessed by children should acknowledge relevant complaints for all services and complaints for services likely to be accessed by children with indicative timeframes for resolution and information about the resolution of complaints.
d) **Measures UR4:** Providers of all U2U services likely to be accessed by children should take appropriate action in response to complaints for services likely to be accessed by children.
e) **Measure UR5:** Providers of all search services likely to be accessed by children should take appropriate action in response to complaints for services likely to be accessed by children.

18.20 Although we assess the impact of these measures separately below, they interrelate in a number of instances and should be considered as a package. For example, Measure UR3 relates to the duty to take appropriate action in response to complaints, like Measures UR4 and UR5, as well as to the duty to operate transparent complaints processes, like part of Measure UR2. Furthermore, much of the evidence we rely on to support these measures is cross-cutting. Where this is the case, we refer back to relevant evidence, rather than repeating it in multiple places in the Section.

[437] See Sections 21(5) and 32(5) of the Act.

Measure UR1: Have complaints processes which enable people to make relevant complaints for services likely to be accessed by children

Explanation of the measure

18.21 Sections 21(2)(a) and 32(2)(a) of the Act require that all providers of U2U and search services provide a means for people to submit relevant kinds of complaints.

18.22 The Act sets out that for providers of **U2U services** likely to be accessed by children relevant complaints are:[438]

a) Complaints by users and affected persons about content, present on a part of the service that children can access, which they consider to be content that is harmful to children.
b) Complaints by users and affected persons if they consider that the provider is not complying with the safety duties protecting children.
c) Appeals by users whose content may have been incorrectly identified as being content harmful to children, leading to it being removed or access to it restricted.
d) Appeals by users who have received a warning or whose accounts have been suspended or banned or otherwise restricted for generating, uploading or sharing content the provider of the service considers to be content harmful to children.
e) Complaints by a user who is unable to access content because measures used to comply with the safety duties protecting children have resulted in an incorrect assessment of the user's age.

18.23 For providers of **search services** likely to be accessed by children, relevant complaints are:

a) Complaints by users and affected persons about search content (i.e., the results of a user's search query) which they consider to be content that is harmful to children.
b) Complaints by users and affected persons about a service not complying with the safety duties protecting children.
c) Appeals by a website owner whose website or database may have been incorrectly identified as containing content harmful to children, leading to it "no longer appearing in search results or being given a lower priority in search results".
d) Complaints by a user who is unable to access content because measures used to comply with the safety duties protecting children have resulted in an incorrect assessment of the user's age.[439]

18.24 Under the Act, to be a "user", it does not matter whether the person is registered to use a service.[440] Therefore, all users and affected persons must be able to make relevant complaints for services likely to be accessed by children, regardless of whether they are registered with the service or logged into the service. We understand that this is currently common practice for many services. For example, in response to our 2022 Illegal Harms Call for Evidence ('our 2022 CFE'), Google told us that it is not necessary for a user to create an account to report content on YouTube.[441] Likewise, Pinterest has a reporting form on their

[438] Section 21(5) of the Act.
[439] Section 32(5) of the Act.
[440] See section 227 (2) of the Act.
[441] Google response to 2022 Illegal Harms Call for Evidence.

website that can be accessed by anyone, whether they are a Pinterest user or not.[442] TikTok users similarly do not need to be logged in to flag content.[443] This suggests that it is technically feasible to create complaints processes open to all users and affected persons.

18.25 The Act only requires service providers to accept relevant complaints for services likely to be accessed by children from users, affected persons and website owners in the UK. This means that to comply with the Act, either service providers need to be able to recognise relevant complaints for services likely to be accessed by children from complainants in the UK, or they need to handle all complaints as though they were relevant complaints for services likely to be accessed by children from complainants in the UK.

18.26 In order to do the former, for complaints about content a user or affected person considers to be content harmful to children, service providers may want to establish a way for users and affected persons to tell them whether the content they are complaining about was served to them in the UK. Complaints about content that was not served in the UK would not be relevant complaints for services likely to be accessed by children.

18.27 For appeals, service providers may want to inform users and website owners when they are entitled to submit an appeal under the Act because action has been taken about their content, account or website for being content harmful to children; or they may want to establish a system that allows them to recognise when an appeal is a relevant complaint that should be handled in accordance with the Act.

18.28 Search providers may also need a way for complainants to tell them if they are a website owner, because only website owners have a right to make certain kinds of complaints.

18.29 We therefore propose to recommend that providers of all U2U and search services likely to be accessed by children have complaints processes that enable UK users, affected persons and (for search services where relevant) interested persons to make each type of relevant complaint in a way which will ensure that the service provider will take appropriate action in response to them. We consider this the minimum necessary to comply with the Act.

18.30 This measure mirrors an equivalent one in the draft Illegal Content Codes, which recommends all service providers operate complaints processes that enable people to make other types of complaints, such as complaints about suspected illegal content. Providers who should apply both measures may operate a single complaints process for various different types of complaints, if they wish to do so.[444]

Rights assessment

18.31 This proposed measure recommends that providers of services within scope have complaints processes that enable users and affected persons, and for search services also interested persons, to make complaints about relevant matters set out in the Act.[445]

18.32 The Act requires that providers of all U2U and search services that are likely to be accessed by children, operate a complaints procedure that allows for a variety of types of complaints

[442] Pinterest response to 2023 Protection of Children Call for Evidence.
[443] Ofcom, 2022. Ofcom's first year of video-sharing platform regulation.
[444] See our Illegal Harms Consultation, Volume 4, Section 16 for discussion of why providers cannot accept all complaints through their content reporting tool.
[445] Which includes reports of content that is either harmful to children or is not permitted by the service's terms.

and appropriate action to be taken in response.[446] We have proposed similar measures in our Illegal Harms Consultation.

18.33 As a result of a complaint, service providers may take steps that affect the rights of users and others who have raised complaints (including both children and adults) to privacy (Article 8 of the ECHR), freedom of religion and belief (Article 9 of the ECHR), freedom of expression (Article 10 of the ECHR) and freedom of association (Article 11 of the ECHR). We have therefore considered the extent to which the degree of interference with these rights is proportionate.[447]

Freedom of expression and association

18.34 As explained in Section 2, Scope of this consultation, Article 10 of the ECHR upholds the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 11 of the ECHR upholds the right to associate with others. The right to freedom of expression and association are qualified rights. Ofcom must exercise its duties under the Act in light of users', affected or interested persons' and service providers' Article 10 (and Article 11) rights and not interfere with that right unless it is satisfied that it is necessary and proportionate to do so.

18.35 Users also include those who are operating on behalf of a business, or accounts that might also be concerned with other entities, such as charities, as well as those with their own, individual account. Both corporate and individual users can benefit from the right to freedom of expression, and we acknowledge the potential risk of interference with the rights of these users to freedom of expression, in addition to the rights of children and adults as individuals. For ease of reference, when we refer to rights of adult users, we include those who are acting on behalf of a business or other entity.

18.36 With this proposed measure, potential interference with both child and adult users' rights to freedom of expression and association may arise where the service provider decides, for example as a result of a complaint, to restrict access to material it considers to be harmful to children, or restricts users' ability to use the service (i.e. banning or suspending them), on the basis of incorrect assessments of the nature of the content. Restrictions may also arise if users' ability to access the service (or part of it, including any content deemed to be harmful to children) due to an incorrect assessment of a user's age (e.g., through an age assurance process). We consider that the impact on users (including children and adults) can be significantly mitigated by having a mechanism for appealing against incorrect decisions.

18.37 We are also provisionally of the view that this proposed measure could have positive impacts on the rights of users (including adults and children) to freedom of expression and freedom of association. For example, a process for raising complaints with the service about content harmful to children could result in more effective content moderation creating safer spaces online where children may feel more able to join online communities and receive and impart (non-harmful) ideas and information with other users. This measure could therefore also have significant benefits to children, in terms of safeguarding their rights to freedom of expression and assembly in safer online spaces, as well as in terms of protecting them from exposure to harm. Adult users would also benefit from this proposed measure if, for example their age is incorrectly assessed, by enabling them to alert the service of the error

[446] Sections 21(2) and 32(2) of the Act.
[447] Including affected or interested persons as appropriate depending on the type of service provider.

and restore appropriate access to legal content and online communities that restrict children from accessing.

18.38 We therefore consider that the impact of the proposed measure as a result of services' complaints and reporting processes on child and adult users' rights to freedom of expression, above and beyond the requirements of the Act, to be relatively limited, and is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that the proposed measure is therefore proportionate.

18.39 The proposed measure may also have an impact on service providers' rights to freedom of expression as, to the extent that they do not already operate a complaints procedure that provides for all relevant complaints set out in the Act, providers would need to put in place steps to ensure that they have this in place. However, most of this impact arises from the duties placed on service providers under the Act by the UK Parliament, and we are allowing flexibility for providers as to the precise approach and action they take to secure the outcomes required by the duties. We therefore consider that to the extent that the proposed measure affects service providers' rights to freedom of expression, it is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

Privacy

18.40 As explained in Section 2, Scope of this consultation, Article 8 of the ECHR confers the right to respect for individuals' private and family life. An interference with the right to privacy must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest. Again, in order to be 'necessary', the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.

18.41 All complaints processes will involve the processing of personal data of individuals, including children and those who are not users of the service, such as affected or interested persons. It will therefore affect users' rights to privacy and their rights under data protection law. The degree of interference will depend to a degree on the extent to which the nature of any affected content and communications is public or private, or, in other words, gives rise to a legitimate expectation of privacy. This proposed measure is not limited only to content or communications that are communicated publicly and may lead to the review of content or communications in relation to which individuals might expect a reasonable degree of privacy, which would in turn lead to more significant privacy impacts than in connection with impacts on content and communications that are widely publicly available (whether on the service concerned or more generally).[448] The impact on users' or other individuals' rights would also be affected by the nature of the action taken as a result of the complaints

[448] As part of its Illegal Harms Consultation Ofcom consulted on draft guidance on content communicated 'publicly' and 'privately' under the Act. That guidance recognises that whether content is communicated 'publicly' or 'privately' for the purposes of the Act will not necessarily align with whether that content engages users' (or other individuals') rights to privacy under Article 8 of the European Convention on Human Rights. For example, it is possible that users might have a right to privacy under Article 8 of the ECHR in relation to content which is communicated 'publicly' for the purposes of the Act. Conversely, users may not have a right to privacy under Article 8 of the ECHR in relation to content which is nevertheless communicated 'privately' for the purposes of the Act.

process. For example, the level of intrusion and significance of the impact is likely to be higher where the outcome of the complaint is judged to warrant restrictive measures are applied such as banning or suspending an individual, compared to less restrictive measures such as downranking or age restricting content.

18.42 The duty for service providers to operate a complaints procedure that enables relevant complaints is a requirement of the Act, and not of this proposed measure, and we are giving service providers flexibility as to precisely how they implement this and what action they take. We recognise that depending on how service providers decide to implement the proposed measure, it could result in a greater or lesser impact on users' privacy rights. However, as noted above, it remains open to service providers (and in the exercise of their own rights to freedom of expression) to decide how to operate their complaints procedure, and what forms of personal data they consider they need to gather to process complaints, so long as they comply with the Act and the requirements of data protection legislation.

18.43 We acknowledge the potential risk of negative impacts on the right to privacy, for example where content is categorised as harmful to children incorrectly, or where the age of a user is incorrectly assessed. The degree of impact will also depend on the extent of personal data about individuals which may need to be processed in order to review and respond to a complaint. The proposed measure does not specify that service providers should obtain or retain any specific types of personal data about individual users, and we consider that service providers can implement the measure in a way which minimises the amount of personal data which may be processed or retained so that it is no more than needed to handle and respond appropriately to the complaint. In processing users' personal data for the purposes of this measure, service providers would need to comply with relevant data protection legislation. This means they should apply appropriate safeguards to protect the rights of both children, whose personal data may require special consideration, and adults.[449] When implementing complaints procedures, service providers should have regard to the ICO Commissioner's Opinion on Age Assurance for the Children's code, and comply with the standards set out in the ICO's Age Appropriate Design Code in respect of children's personal data, along with other relevant guidance from the ICO.[450] Providers may also use third parties to carry out complaints processes on their behalf and ICO guidance is clear that service providers should ensure that individuals' rights to privacy are fully protected when a third party has access to their personal data. [451]

18.44 We therefore consider that the impact of the proposed measure as a result of service providers' implementation of a complaints procedure on child and adult users' rights to privacy to be relatively limited, and (assuming service providers also comply with data protection legislation requirements) it is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

[449] In line with Recital 38 UK GDPR.
[450] ICO.org.uk. Information Commissioner's Opinion. [accessed 19 April 2024]. ICO.org.uk. Age Appropriate Design Code. [accessed 19 April 2024]. UK GDPR Guidance. [accessed 19 April 2024]
[451] Further information on the requirements for contracts between data controllers and processors can be found at Contract and liabilities between controllers and processors. [accessed 19 April 2024].

Impacts on services

18.45 Handling relevant complaints for services likely to be accessed by children is required by the Act. Given that our proposed recommendation closely follows the specific requirements in the Act and leaves the widest possible discretion to providers on how to achieve what is required, we consider its impacts are required by the Act.

18.46 Providers can decide the most appropriate and proportionate approach for their own contexts, and the set-up and ongoing costs that flow from that are imposed by the Act. Many service providers already allow user complaints, and so incremental costs are expected to be minimal and relate to staff time to ensure complaints processes are fit for purpose. Costs will be higher for service providers that may not currently allow user complaints in any form.[452] This flexibility will allow them to take an approach proportionate to the risks they carry.

18.47 All service providers who should apply this measure should also apply the related proposed measure in our Illegal Harms Consultation.[453] We consider that there are likely to be substantial overlaps in costs between the two measures because we expect providers are likely to use the same complaints processes to accept different types of complaints. Service providers that allow users to categorise complaints may incur a minimal additional cost to add categories covering content harmful to children.

Which providers we propose should implement this measure

18.48 As discussed above, this measure codifies the requirement in the Act for service providers to accept relevant complaints for services likely to be accessed by children, and we consider this is the minimum necessary to comply with the Act. We therefore have provisionally concluded to recommend this measure to all U2U and search services likely to be accessed by children.

Other options considered

18.49 Two respondents to our 2023 Protection of Children Call for Evidence ('our 2023 CFE') called for reporting and complaints procedures to be standardised across services to make it easier for users to submit a report or complaint.[454] While we recognise that there may be some benefits from this, we had concerns that, given the wide range of services in scope of the Act, to impose a standardised process for all service providers, without regard for their size, user base, service characteristics, risk profile, or the nature of the content they host, would be unlikely to be an effective or proportionate solution.[455] It might also restrict the ability of service providers to adopt innovative approaches to receiving and handling complaints.

[452] Department for Digital, Culture, Media and Sport, 2022. Online Safety Bill: Impact assessment. [accessed 19 April 2024].
[453] Our Illegal Harms Consultation, Volume 4, Section 16, Measure 1.
[454] Samaritans response to 2023 Protection of Children Call for Evidence. Resolver, a Kroll business (formerly Crisp) response to 2023 Protection of Children Call for Evidence.
[455] Children participating in our 2024 qualitative research into children's attitudes to reporting content online said they thought consistency across services would make it easier to submit reports about harmful content. Ofcom, 2024. Children's Attitudes to Reporting Online Content.

Provisional conclusion

18.50 Given the harms this measure seeks to mitigate in respect of content harmful to children, as well as service providers duties to provide a means for people to make complaints about content harmful to children and other types of complaints, we consider this measure appropriate and proportionate to recommend for inclusion in the draft Children's Safety Codes. For the draft legal text for this measure, please see PCU C1 in Annex A7 and PCS C1 in Annex A8.

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

Explanation of the measure

18.51 Sections 20(2) and 31(2) of the Act place duties on providers of U2U and search services to operate systems and processes that allow people in the UK to easily report content harmful to children. Complaints processes for all types of relevant complaint must be easy to access, easy to use (including by children) and transparent (see section 21(2)(c) and 32(2)(c) of the Act).

18.52 As discussed in the 'Interaction with Illegal Harms' section above, as part of this measure we are making proposals that mirror equivalent proposals in the draft Illegal Harms Codes. We are also proposing to add an additional element to this measure to protect children from harm in light of new evidence regarding the barriers children face to reporting. We discuss the equivalent proposals and then the additional element in turn below.

Equivalent proposals

18.53 In our Illegal Harms Consultation we set out evidence that suggests complainants may find it difficult to make complaints for a number of reasons. As we discuss in the 'Reasons why complaints processes are underused' section above, and in Section 7.11, Governance, systems and processes, evidence suggests that children specifically currently face a number of barriers to complaining.[456] This implies that providers may not be doing enough currently to ensure their processes are easy to find, easy to access and easy to use for all users. We made recommendations to reduce these barriers in Section 16, Reporting and complaints, of our Illegal Harms Consultation. We are now provisionally proposing to extend those recommendations to apply to the additional types of complaints which providers of services likely to be accessed by children are required to handle.

18.54 We are proposing to recommend that providers of all U2U and search services likely to be accessed by children should provide an easy to find, easy to access and easy to use complaints process including:

a) **Measure UR2 (a):** Having easily findable and accessible content reporting tools and ways to make other complaints;
b) **Measure UR2 (b):** Ensuring information and processes relating to complaints are accessible and comprehensible, with services having regard to the needs of their userbase, including children;

[456] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.

c) **Measure UR2 (c):** Having as few steps as reasonably practicable to make a complaint; and
d) **Measure UR2 (d):** Enabling complainants to include context/supporting material when making a complaint.

18.55 We provisionally consider that proposed measures UR2 (a), (b), (c) and (d) are the minimum necessary to ensure complaints procedures are easy to find, easy to access and easy to use as required by the Act. We therefore consider them proportionate for providers of all U2U and search services likely to be accessed by children. We discuss these recommendations in the 'Equivalent proposals' section below.

Additional element

18.56 Since publishing our Illegal Harms Consultation, we have become aware of new evidence regarding children's concerns about confidentiality of complaints processes on U2U services and whether the person whose content they complain about will find out who made the complaint.[457] This evidence suggests that many children are unclear about what happens following submission of a complaint and that providers may not currently be doing enough to ensure their complaints procedures are transparent. In light of this evidence, which we discuss below, we propose to recommend an additional measure for providers of U2U services which was not included in our draft Illegal Content Codes. The evidence for this measure does not relate to specific harms or types of content, but rather children's concerns about complaints in general. We consider that this measure will help to make complaints processes more transparent in relation to all types of complaints. We are therefore proposing to include it in both the draft Illegal Content Codes and the draft Children's Safety Codes.

18.57 We are proposing to recommend that providers of all U2U services likely to be accessed by children should:

a) **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 (and what information is provided to the complained about user regarding the original complaint and complainant if they subsequently appeal).

18.58 As we explain below, we are recommending this measure for providers of all U2U services likely to be accessed by children in relation to complaints about suspected illegal content and about content considered content harmful to children.

18.59 We discuss the evidence, costs, and impacts of this measure in the 'Additional element' section below.

[457] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.

Equivalent proposals: Measures UR2 (a), (b), (c) and (d)

Measure UR2 (a): tools for reporting content harmful to children should be easy to find and easily accessible in relation to the content being viewed; and processes for making other types of complaints should also be easy to find and easily accessible.

Measure UR2 (b): information and processes relating to complaints should be accessible and comprehensible, including to children; and services should have regard to the findings of their risk assessment concerning the accessibility needs of their UK user base.

Measure UR2 (c): the number of steps necessary (such as the number of clicks or navigation points) for people to submit any complaint should be as few as is reasonably practicable.

Measure UR2 (d): complainants should be able to provide relevant information or supporting material when submitting complaints to a service.

Effectiveness at addressing risks to children

18.60 We set out below evidence for the effectiveness of our proposals at reducing the barriers children face to complaining and enabling providers to ensure their complaints processes are easy to access and easy to use. This in turn will help ensure services providers' complaints processes are effective for protecting children from content harmful to children present on their services.

18.61 We currently have less evidence relating to how easy users, affected persons and website owners find it to complain to providers of search services. For this reason, much of the evidence referred to below relates to U2U services. However, we provisionally consider that most of it can be extrapolated to apply to search services, since the principles of what complainants will find easy or difficult in a complaints process will be the same regardless of the type of service.

18.62 The analysis below does not differ substantially from the discussion in Section 16, Reporting and complaints, of our Illegal Harms Consultation. However, we have added new evidence where appropriate, and reordered our recommendations to make our rationale easier to follow. We have also updated our language in places to clarify our recommendations.

Measure UR2 (a): reporting tools and complaints processes should be easy to find and access.

18.63 The research discussed in the 'Reasons why complaints processes are underused' section above, and in Section 7.11, Governance, systems and processes, suggests that where content reporting tools are hard to find this can discourage or prevent users (and especially children) from reporting content, including content considered harmful to children.[458] We consider that in order for complaints processes to be easy to access, as required by the Act, it is vital for providers to reduce this barrier, so that users are able to easily locate those reporting tools and alert providers to potentially harmful content on their services.

18.64 We know that it is currently common practice on several large services for reporting tools to be placed behind other buttons or features. In response to our 2022 and 2023 CFEs, Meta told us that on their services, Facebook and Instagram, the reporting tools are behind the 'three dots' feature, while on WhatsApp users have to long press a message to access a drop down menu with an option to report.[459] In the TikTok app, users can also access the

[458] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.
[459] Meta's response to 2022 Illegal Harms Call for Evidence. WhatsApp's response to 2023 Protection of Children Call for Evidence.

reporting tool by long pressing a video, or they can find it behind the 'share' button.[460] We know from our own desk research that similarly on YouTube the report tool is located behind the 'three dots' icon (desktop) or the settings symbol (app).[461] Users should press and hold down content they wish to report on Snapchat.[462] On Google Search, users can either click on 'three dots' that appear alongside search results or go into their settings to report search results.[463] On Bing, users can report search results on the Microsoft website, but not alongside search results themselves.[464]

18.65 Our research suggests that making the reporting tool more prominent can make it easier to access and increase the number of reports users make. Our behavioural research into different designs for content-reporting tools on VSPs found that inserting a 'flag' icon on the main options bar, rather than including it in a drop-down menu behind an ellipsis, led to four times more reports by adult users.[465] Moving the reporting tool to the main options bar also had the effect of reducing the number of steps needed to submit a complaint. We discuss this further under 'Measure UR2 (c)' below.

18.66 Although this trial was conducted with adults, we think it is likely that improving the prominence of the reporting tool will also make it easier for children to report. Many respondents to our 2023 CFE recommended reporting tools should be prominent and clearly identifiable to make it easier for children to find them.[466]

18.67 Other respondents to our CFEs also suggested that reporting tools should be prominent and easy to find. The Center for Countering Digital Hate said in its response to our 2022 CFE that it found in some cases it was difficult for users to find a reporting tool and recommended that "platforms can improve how to find a reporting function by adopting a safety by design approach".[467] The LEGO Group suggested that 'ensuring reporting and complaints links/functions are well placed within a digital experience … increase[s] the likelihood of young users utilizing the services. This includes contextual placement, so that users have access to reporting functions at the point where they are most likely to need it'.[468] Other respondents suggested providers should offer easy to use reporting tools, such as clickable buttons, or that tools should be made clear to users by using an easily recognisable symbol, such as a flag, or words like 'Report'.[469] Participants in our research on reporting behaviours and attitudes in children said reporting could be simplified through the report button being more visible, being placed in a consistent spot like the top corner and separating it from other buttons. Participants also said that bright colours can make reporting tools easier to find, for example a red flag button.[470]

---

[460] Ofcom, 2022. Ofcom's first year of video-sharing platform regulation.
[461] Ofcom desk research, conducted March 2024.
[462] Snap, How do I report abuse or illegal content on Snapchat?. [accessed 26 February 2024].
[463] Google, Report content on Google. [accessed 14 March 2024].
[464] Bing, Report a concern to Bing. [accessed 14 March 2024].
[465] Ofcom, 2023. Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour.
[466] Girlguiding response to 2023 Protection of Children Call for Evidence; 5Rights response to 2023 Protection of Children Call for Evidence; UKSIC response to 2023 Protection of Children Call for Evidence; Ruth Moss response to 2023 Protection of Children Call for Evidence.
[467] Center for Countering Digital Hate response to 2022 Illegal Harms Call for Evidence.
[468] LEGO Group response to 2022 Illegal Harms Call for Evidence.
[469] Catherine Knibbs Ltd – trading as Children and Tech response to 2022 Illegal Harms Call for Evidence; TrustElevate response to 2022 Illegal Harms Call for Evidence.
[470] Ofcom, 2024. Children's Attitudes to Reporting Content Online.

18.68 To reduce barriers to reporting, we therefore propose to recommend that reporting tools should be easy to find and easily accessible in relation to the content being viewed.

18.69 We have less evidence that it is difficult for complainants to make other kinds of relevant complaints for services likely to be accessed by children, such as appeals. However, the Act also requires those kinds of complaints procedures to be "easy to access". We do not consider that a complaints procedure is easy to access if it is not clear to users where and how they can make a complaint. We therefore consider that processes for making other kinds of complaints should also be easy to find and easily accessible.

18.70 Given the wide range of services who should apply this measure, and the wide range of user interfaces they may adopt, we do not think it is appropriate to be prescriptive about where exactly complaints tools should be located or what they should look like. However, we would encourage providers to take note of the evidence referred to here, which consistently suggests that clearly identifiable symbols, located close to the content being viewed can make complaints tools easier to find.

Measure UR2 (b): information and processes relating to complaints should be accessible and comprehensible.

18.71 For complaints tools to be easy to access and easy to use, including by children, we also consider that, as a minimum, information and processes relating to complaints should be accessible and comprehensible.[471]

18.72 Evidence suggests that one way service providers could ensure their processes relating to complaints are accessible and comprehensible to children would be to make instructions on how to complain easy for children to find prior to making a complaint. As discussed in Section 7.11, Governance, systems and processes, research suggests that some children are put off complaining because they do not know how to complain or believe it will be difficult.[472]

18.73 This view was echoed by many respondents to our 2023 CFE, who recommended that children should be provided with clearer information and guidance on how to report.[473] In their response, Samaritans said that services should provide "step-by-step information on how to report and what actions may be taken. This information should be clearly displayed to new users, and existing users regularly reminded".[474] NCMEC similarly suggested that "safety messaging [in the terms of service] should include clear guidance on how to report troubling content and users".[475]

18.74 In our 2024 research into children's attitudes to reporting, many children said that if they were unsure of how to submit a report, and wanted more information, Google Search would be the first place they would look. Some participants said they would ask for advice from an authority figure such as a parent, teacher, or group administrator on the service. A few also

[471] The Act contains specific requirements, considered in Section 19, about how complaints processes should be described in U2U and search services' terms of service and publicly available statements respectively. This measure does not relate to those duties.

[472] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.

[473] 5Rights response to 2022 Illegal Harms Call for Evidence; Nexus response to 2023 Protection of Children Call for Evidence.

[474] Samaritans response to 2023 Protection of Children Call for Evidence.

[475] NCMEC response to 2023 Protection of Children Call for Evidence.

mentioned they would look in the on-service help section. [476] These findings suggest that providers do not always make it as easy as they could for children to access support about how to submit a report, instead relying on children to proactively search for guidance.

18.75 Evidence also indicates that providers could make their processes relating to complaints accessible and comprehensible to children by making an explanation of the actions they may take in response to complaints easy to find prior to making a complaint.

18.76 Research suggests that many children do not currently understand what happens in response to complaints, with many believing services take no action.[477] Being clear with children about what actions service providers may take in response to complaints could help to increase children's trust in complaints processes and make complaints processes comprehensible to children, thereby reducing barriers to complaining. Several respondents to our 2023 CFE echoed this view. In their responses, 5Rights, Samaritans and NCMEC all called for services to explain to users what actions may be taken in response to a complaint.[478]

18.77 We know that some services already make this information available to users. In their response to our 2022 CFE, Google told us that YouTube has produced a video on 'the life of a flag' to help users understand what happens to content they have flagged.[479] This suggests that while there may be a wide range of actions providers may take in response to complaints, it is still possible to give an indication of these for the purposes of transparency.

18.78 Terms and Statement measure TS1 in Section 19, recommends that service providers should explain in their terms of service (U2U services) or publicly available statements (search services) the processes and policies that govern the handling and resolution of relevant complaints for services likely to be accessed by children. User support measure US6 in Section 21, recommends that providers of services that are multi-risk for content harmful to children should also provide age-appropriate support materials for children and parents explaining how to use reporting tools. These measures will help providers to meet their duty in the Act to make this information easily accessible, including to children.[480] As part of implementing those recommendations, providers could include instructions on how to complain and an explanation of the actions the provider may take in response to complaints. This would help to ensure the whole complaints process is accessible and comprehensible, including to children.

18.79 We recognise that what people find accessible and comprehensible will vary person-to-person depending on abilities and needs. We understand that bespoke systems for all types of users, affected persons and website owners with vulnerabilities may not be feasible. However, providers should aim to ensure complaints procedures are accessible for as many people as practically possible. Some groups, such as children, and people with certain disabilities, may have particular accessibility needs that providers should take into account when designing their processes.

18.80 To take those needs into account, we propose to recommend that providers should have regard to the findings of their risk assessments concerning the accessibility needs of their UK

[476] Ofcom, 2024. Children's Attitudes to Reporting Content Online.
[477] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.
[478] Samaritans response to 2023 Protection of Children Call for Evidence; NCMEC response to 2023 Protection of Children Call for Evidence; 5Rights response to 2023 Protection of Children Call for Evidence.
[479] Google response to 2022 Illegal Harms Call for Evidence.
[480] See sections 21(3) and 32(3) of the Act.

user base, including children and people with disabilities, when designing their complaints procedures. We consider this would also be likely to make it easier for all users, affected persons and website owners to make complaints.

18.81 In practice, we consider that as a minimum this means that written information should be comprehensible based on the likely reading age of the youngest person permitted to access the service without the consent of a parent or carer (for example, as set out in the service's terms of service). The process should also be designed with accessibility in mind and for the purposes of ensuring the information can be used by those dependent on assistive technologies including keyboard navigation, and screen reading technology.[481] We set out similar minimum standards for terms of service and publicly available statements in measure TS2 of Section 19. We know aspects of this are already current practice on certain services. For example, in its response to our 2022 CFE, Google told us that it has a 'read aloud' option for the information in its help centre about how to report, to make it accessible to a wider range of people.[482]

Measure UR2 (c): The number of steps to submit a complaint should be as few as reasonably practicable.

18.82 The research cited in the 'Reasons why complaints processes are underused' section above, and discussed in Section 7.11, Governance, systems and processes, also finds that children are discouraged from reporting because complaints processes are perceived to be time consuming and burdensome.[483] We provisionally consider that in order for complaints procedures to be easy to use, as required by the Act, providers should ensure that submitting a complaint is as quick and straightforward as possible.

18.83 Evidence suggests that reducing the number of steps needed to submit a complaint would help achieve this. In the behavioural research trial discussed at under 'Measure UR2 (a)' above, adding a flag icon to the main menu, rather than including it in a drop-down menu behind an ellipsis, also had the effect of reducing the number of steps in the process. As mentioned above, this change led to a statistically significant increase in the number of adult users reporting content they were concerned about.[484] This evidence suggests that decreasing the number of steps needed to submit a complaint could make complaints processes easier to use. Further to this, participants in our 2024 qualitative research into children's attitudes to reporting said they found short, clear reporting processes, with as few steps as possible, appealing.[485]

18.84 This was also raised by respondents to our CFEs. In response to our 2022 CFE, TrustElevate recommended that "reporting mechanisms should include the minimum number of clicks and steps for a user to quickly submit a report or complaint with ease while equipping the receiving party/platform with sufficient information to assess the report and determine the appropriate response".[486] The LEGO Group similarly recommended that "reporting

[481] See the Web Content Accessibility Guidelines for further guidance on how to make digital services, websites and apps accessible to everyone. UK Government. Web Content Accessibility Guidelines (WCAG). [accessed April 2024].
[482] Google response to 2022 Illegal Harms Call for Evidence.
[483] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.
[484] Ofcom, 2023. Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour.
[485] Ofcom, 2024. Children's Attitudes to Reporting Content Online.
[486] TrustElevate response to 2022 Illegal Harms Call for Evidence.

processes should be simplified for young users – removing number of pages they pass through to lodge complaint/report."[487]

18.85 In order to help reduce barriers to complaining we propose to recommend that the number of steps necessary to submit a complaint should be as few as reasonably practicable. We think this could make it easier for all users to complain, not just children.

18.86 Given the large number and variety of different types of service in scope of the Act, we provisionally consider it would not be appropriate to set out the maximum number of steps required to make a report or another complaint. However, when considering how to implement this measure, we encourage providers to take note of the findings of our research into children's attitudes to reporting, which found that young people are more likely to use reporting tools when the process only requires a couple of clicks, is straightforward and intuitive to find. Additional steps and complicated reporting flows create barriers.[488]

Measure UR2 (d): complainants should be able to provide supporting material when complaining.

18.87 Regardless of the number of steps involved in submitting a complaint, evidence discussed below suggests that complainants only find complaints processes easy to use, as required by the Act, if the processes allow them to include all the information relevant to the subject of the complaint. For this reason, we provisionally consider that complainants should be able to provide relevant supporting material when submitting a complaint.

18.88 Evidence suggests that this is particularly helpful when complaining about highly contextual harms, like bullying. Some participants in our research into children's experiences of cyberbullying said they found free text boxes where they could provide additional information, in addition to categories helpful when reporting bullying content. [489] We know that the option of free text is already current practice on certain services, including, for example, Twitch.[490] Participants also valued it when services, such as WhatsApp, automatically include recent messages between users in the report, as this makes it quicker and easier to report multiple instances of bullying.[491] Further to this, our 2024 research into children's attitudes to reporting found that children appreciate being able to add further information to contextualise, add detail and/or explain the reason for their report. [492]

18.89 Context is often crucial to enable content moderators to correctly identify content harmful to children. Respondents to our 2022 CFE cited examples of when being able to provide context had helped services identify and remove harmful content. For example, the Antisemitism Policy Trust described an occasion when it had provided additional context when reporting a picture that was being used by far-right actors to intimidate and harass a high-profile Jewish individual. Without additional context, the photo was deemed not to breach the service's rules.[493] Refuge provided an example of survivors of domestic abuse who had received images of their front doors and road signs after moving to a new location. The image of a front door is not harmful in itself so is unlikely to be removed by content

[487] LEGO Group response to 2022 Illegal Harms Call for Evidence.
[488] Ofcom, 2024. Children's Attitudes to Reporting Content Online.
[489] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.
[490] Ofcom, 2022. Ofcom's first year of video-sharing platform regulation.
[491] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.
[492] Ofcom, 2024. Children's Attitudes to Reporting Content Online.
[493] Antisemitism Policy Trust response to 2022 Illegal Harms Call for Evidence.

moderators. However, with added context it may be reasonable to infer that the content amounts to harassment.[494]

18.90 Refuge also said that 'survivors must usually report individual pieces of content in turn and are not able to report a user. Perpetrators will often send dozens or hundreds of messages, making reporting time-consuming and potentially re-traumatising process for survivors.'[495] An ability to provide context, for example, screenshots showing how the user is being subjected to a pattern of behaviour or the identities of the accounts engaging in the behaviour concerned, would reduce this burden on users wishing to raise a complaint.

18.91 Some of the content mentioned in these examples may be illegal content, rather than content harmful to children as defined in the Act. However, we consider that enabling users to provide additional context is also likely to reduce the burden on users when reporting bullying or other abusive content. It would also make it easier for moderators to reach decisions about these types of content, and other content harmful to children.

18.92 If users are unable to provide supporting information when making complaints, services may not have the necessary information to make informed judgements and therefore may decide not to uphold valid complaints. There may also be a risk that people will consider the reporting process difficult to use as a result – particularly if they have to complain multiple times to get content removed – and so decide not to complain at all in the future. This could lead to providers not taking steps to protect children from harmful content, or not being made aware of that content at all, and therefore being unable to take steps to protect children from it or prevent them from encountering it.

18.93 It could also lead to other negative consequences, for example, if complainants cannot provide relevant information to contest a decision about a restriction placed on their content or account. Complainants wishing to complain about a problem with a service's assessment of their age or some form of non-compliance with the safety duty may also need to be able to attach a screen shot or a description to their complaints.

18.94 In light of this evidence, we therefore propose to recommend that service providers should enable complainants to include supporting material when submitting complaints about content harmful to children and when making other types of complaints.

18.95 We do not think it is appropriate at this stage to be prescriptive about what formats of supporting evidence service providers should accept alongside complaints, as this is likely to vary depending on the nature of the service and its userbase. Examples of supporting evidence could include text, screenshots of messages or links to content.

Rights assessment

18.96 This proposed measure recommends that providers of all services in scope operate their complaints procedures in a way that makes them transparent, easy to access and easy to use. We have designed this in a way that allows service providers the flexibility to decide the details of how they achieve this, whilst setting expectations about what that should entail.

Freedom of expression and association

18.97 We consider that this proposed measure has the potential to affect users' and others' (including adults' and children's) rights to freedom of expression and to freedom of

[494] Refuge response to 2022 Illegal Harms Call for Evidence.
[495] Refuge response to 2022 Illegal Harms Call for Evidence.

association, and service providers' rights to freedom of expression, for the reasons set out in relation to Measure UR1 above. We also consider the likely degree of interference with these rights to be limited for the reasons set out in relation to Measure UR1 above.

18.98 In addition to the impacts identified in Measure UR1, we have considered whether there may be a risk that this measure could lead to an increase in false or malicious complaints as a result of increased ease of use and transparency, and whether such an outcome could lead to any additional restrictions on rights to freedom of expression or association. However, we do not consider this would be the case. This is because any such complaints would need to be assessed in the same way as other complaints via the providers' content moderation process. We do not consider that false or malicious complaints would be more likely to be wrongly upheld than any other type of complaint, instead we would expect them not to be upheld if providers' content moderation systems were working effectively. We also consider that the Content Moderation Measures (outlined in Section 16) and Search Moderation Measures (outlined in Section 17) would, if followed, reduce the likelihood this would occur.

18.99 As under Measure UR1 we consider the proposed measure could also have positive impacts on freedom of expression and freedom of association rights of children, for example, by enabling children to easily report content harmful to children or complain about a failure to comply with the children's safety duties, could result in safer spaces online where children may feel more able to join online communities and receive and impart (non-harmful) ideas and information with other users. This measure could therefore also have significant benefits to children, in terms of safeguarding their rights to freedom of expression and assembly in safer online spaces, as well as in terms of protecting them from exposure to harm.

18.100 We also consider that there could be positive impacts on the rights of adult users and others, particularly where they have had their access to the service restricted or access to content they have uploaded restricted on the basis that it is content harmful to children. This proposed measure would make it easy to make such complaints and for users to appeal decisions that are incorrect, thus enabling them to exercise their rights more freely and without unnecessary interference.

18.101 We therefore consider that the impact of the proposed measure on users' or others' (including adults' and children's) rights to freedom of expression and of association to be limited and is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that the proposed measure is therefore proportionate.

Privacy

18.102 We consider that this proposed measure – UR2 (a)-(d) – has the potential to affect users' and others' (including adults' and children's) right to privacy for the reasons set out in relation to Measure UR1 above.

18.103 Beyond the impacts identified in Measure UR1, we provisionally consider it unlikely that this proposed measure will create any significant negative impacts on individuals' (including children and adults) rights to privacy.

18.104 However, we are of the view that this proposed measure could have positive impacts on individuals' rights to privacy as it would facilitate complaints by making the process easy to follow and allow individuals to see what outcomes may be applied. This could be especially

important in enabling individuals to exercise their rights in respect of personal data, such as challenging an incorrect assessment of age that results in a restriction of their use of the service.[496]

18.105 As with proposed Measure UR1, service providers are required to comply with data protection laws and to consult relevant guidance from the ICO as required.

18.106 We therefore consider that the impact of the proposed measure as a result of service providers' implementation of a complaints procedure on users' (both adults and children) and others' rights to privacy to be relatively limited, and (assuming service providers also comply with data protection legislation requirements) it is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

Impacts on services

18.107 The recommendations discussed above (Measures UR2 (a), (b), (c) and (d)) describe how we recommend providers meet their duties in the Act relating to the ease of access and ease of use of complaints processes. We are not proposing to specify precisely how providers should design their complaints processes, but instead propose to set out high-level requirements that give providers flexibility to decide how to achieve what is required. This means providers can decide the most appropriate and proportionate approach for their own contexts and risks. The costs of the proposed measure therefore relate to the specific requirements in the Act, over which Ofcom has no discretion.

18.108 Changes to make complaints processes easy to find, easy to access and easy to use may entail some direct one-off costs for designing the required changes. There may also be engineering costs of testing and implementing those changes and further maintaining a complaints process, as providers would need to ensure it continues to meet the requirements over time. These implementation costs would depend on the complexity of the complaints process the provider chooses to adopt. We expect this to vary by service size to some extent, as providers of smaller services will tend to have simpler processes than providers of larger services.

18.109 There will also be ongoing costs of considering complaints. If the complaints process is easier to access and use, the volume of complaints is likely to increase, tending to increase costs. This measure intends to increase the volume of complaints providers receive about content harmful to children, and the costs of dealing with this will tend to increase in proportion to the benefits of the measure.

18.110 Additionally, we believe any potential increase in costs is mitigated by the flexibility allowed in our proposals. In particular, we are not proposing to specify how providers should categorise complaints or exactly what complaints processes should look like. We consider in measures UR4 and UR5 below what appropriate action is in response to complaints.

18.111 We have also considered that complaints may not always accurately identify harmful content.[497] If complaining is easier, it could lead to an increase in the number of complaints

[496] Further guidance on these rights can be found on the ICO website at Individual rights. [accessed 19 April 2024].

[497] For example, TrustPilot's 2021 transparency report says that only 12.4% of consumer user reports in 2021 were deemed to be accurate [accessed December 2023]. Reddit's 2021 transparency report showed that there

about non-harmful content and in the costs of handling such complaints. However, in our research into the effect of making reporting tools more visible and simplifying the process for adults, we observed an increase in the number of reports of harmful content without increasing the number of reports of non-harmful content.[498] While we do not have specific evidence for children's responses to such changes, and are aware that responses may vary by service and with the exact implementation of the change, this could indicate that this measure may not lead to a significant increase in inaccurate reports. User support measure US6 in Section 21 should help to mitigate this risk further by providing age-appropriate resources to children and parents that will help them understand how to report harmful content if they encounter it on a service. Overall, we consider that the costs of handling an increased number of inaccurate reports are likely to be outweighed by the benefits of increased accurate complaints.

18.112 We recognise that all service providers who should apply this measure should also apply the related proposed measure in the Illegal Harms consultation.[499] Measures UR2 (a), (b), (c) and (d) closely mirror measures in our draft Illegal Content Codes, which specifically mentions children as a relevant user group when considering ease of use and ease of access. Providers who adopt the recommendations of the measure in the draft Illegal Content Codes are only expected to incur small additional costs from extending their processes to relevant complaints for services likely to be accessed by children.

Which providers we propose should implement this measure

18.113 Under the Act, providers of all regulated U2U and search services likely to be accessed by children are required to have reporting tools which allow users and affected persons to easily report content that they consider to be content harmful to children. They must also have complaints processes which are easy to access and easy to use, including by children.

18.114 We set out above a number of proposals which we believe, when taken together, would enable providers to meet these duties, as required by the Act. While the measure will have some costs, we provisionally conclude that given the importance of good reporting and complaints procedures, such costs are proportionate for all providers of services likely to be accessed by children and are primarily based on the requirements of the Act.

18.115 This is particularly the case since: (i) it is difficult to envisage how service providers could comply with their duties under the Act if they did not follow the measures that we have set out; and (ii) our approach allows service providers significant flexibility to implement the above measures in a way which is cost effective and practicable for them.

18.116 Therefore, we have provisionally concluded to recommend measures UR2 (a), (b), (c) and (d) to all U2U and search services likely to be accessed by children.

Additional element: Measure UR2 (e)

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 (and what information is

were 31.3m user reports and it acted on 6.27% of these; the rest were duplicate reports, already actioned, or for content which did not violate its rules [accessed December 2023].

[498] Ofcom 2023. Behavioural insights research – understanding the impact of video sharing platform (VSP) design on user behaviour.

[499] Our Illegal Harms Consultation, Volume 4, Section 16, Measure 2.

provided to the complained about user regarding the original complaint and complainant if they subsequently appeal).

Effectiveness at addressing risks to children

18.117 As mentioned above, since publishing our Illegal Harms Consultation, we have become aware of new research which indicates that some children are discouraged from complaining about content on U2U services over concerns about confidentiality and whether the person whose content they complain about will discover who made the complaint.[500] Being more transparent with children about what happens following submission of a complaint could help to dispel this concern and reduce this barrier to complaining.

18.118 Evidence suggests that children would be more likely to complain about content if complaints processes were anonymous. Our 2024 research into children's attitudes to reporting found that concerns about anonymity were a common barrier to reporting amongst children. In particular, participants were concerned about those they reported finding out and causing them harm.[501] Children participating in our research into experiences of cyberbullying said that anonymity was important to them when reporting to ensure that a report or outcome of a report could not be traced back to them, due to concerns that acting against a bully might exacerbate the situation.[502] Similarly, in our 2024 research into children's experiences of violent online content, children said that they did not believe the reporting process would be anonymous, believing that their details would be included in a notification sent to the user they reported. They expressed fear that knowledge of their reporting would eventually become public.[503]

18.119 Thorn's 2021 Responding to Online Threats research into children's attitudes to blocking and reporting found that in a survey of 1,000 children in the USA, 68% said they would be more likely to report if the process was anonymous. This was higher among girls (76% aged 9-12 and 72% aged 13-17).[504] Although Thorn's research primarily focused on 'harmful sexual interactions', we consider that this finding is also likely to be relevant for children's reports of other content and interactions harmful to children as defined in the Act. In their response to our 2022 CFE, the NSPCC cited Thorn's research and called for services to offer reassurances over anonymity and confidentiality as much as possible, to bring about a cultural change towards reporting of harmful content.[505]

18.120 We do not consider it appropriate to recommend at this stage that providers should guarantee that their complaints processes are anonymous (i.e., that no one, including the provider, will be aware of who made the complaint). This is because there may sometimes be legitimate reasons why providers may need to know the identity of the complainant, for example to make a safeguarding or welfare referral. It may sometimes be impossible for providers to prevent users working out that they were complained about and by whom, for instance through a process of elimination or where the content was shared only with one other user.

---

[500] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.
[501] Ofcom, 2024. Children's Attitudes to Reporting Content Online.
[502] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.
[503] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.
[504] Thorn, 2021. Responding to Online Threats: Minors' Perspectives on Disclosing, Reporting and Blocking [accessed November 2023].
[505] NSPCC response to 2022 Illegal Harms Call for Evidence.

18.121 However, we think the evidence suggests that being more transparent with children about what happens when they submit a complaint could help increase trust in complaints processes and make them more likely to complain about content harmful to children. This would help service providers to meet their duties to operate transparent complaints processes and the safety duties protecting children.

18.122 We understand that on many services it is currently common practice not to inform users whose content is removed or restricted how it was detected (i.e., whether it was the subject of a complaint). The evidence discussed above suggests that this is not generally understood by many children. Explaining to children whether users are notified when their content is complained about, and, if so, what information that notification contains (and what information is provided to the complained about user regarding the original complaint and complainant if they subsequently appeal), could therefore help address children's concerns and reduce barriers to complaining.

18.123 Making this information easily accessible could also improve transparency for other users, by informing them of whether they should expect to be notified if their content or account is complained about. We therefore propose to recommend that all providers of U2U services should explain whether they notify users when their content or account is complained about, and if so, what information that notification contains. This explanation should be easily accessible.

18.124 At this stage, we do not have evidence that concerns about confidentiality are a barrier to complaining to providers of search services. We are therefore not proposing to recommend this measure for search services at this time.

18.125 Given the wide range of services who should apply this measure, we do not consider it appropriate to be prescriptive about where exactly this information should be located. However, we consider it should be easily accessible during the complaints process itself. This is because we think this is a key point in the complainant's journey at which this information is likely to be relevant. We understand some services already include information of this nature at this stage of the complaints process. This could be achieved, depending on the type of complaints process, by displaying this information in the complaints form itself (for example, behind a question mark or help button) or including a link to this information alongside the email address to which complaints should be sent (for example, linking to age-appropriate support materials of the kind recommended by User Support Measure US6 of Section 21).

Rights assessment

Freedom of expression and association

18.126 We do not consider that this proposed measure would give rise to any additional restrictions on users' and others' (including adults' and children's) rights to freedom of expression and to freedom of association beyond those already set out in relation to Measure UR1 and Measure UR2 (a)-(d) above. As with Measures UR1 and UR2 (a)-(d) above, we consider that it may, in fact, have positive benefits on these rights and help to safeguard them.

Privacy

18.127 We do consider that this proposed Measure UR2 (e) would give rise to any additional restrictions on users' and others' (including adults and children) right to privacy beyond those already set out in relation to Measure UR1 and Measure UR2 (a)-(d) above. This is because we would not expect the proposed measure to require any personal data to be

processed other than what would already be required to ensure the complaints process works effectively, in line with Measures UR1 and UR2 (a)- (d) above. The notice we are recommending is given in response to a complaint also sets out information that is also likely to be required under data protection laws in respect of the principle of transparency.[506] It will also make clear to individuals what information will be shared with the subject of the complaint, which will help to determine whether individuals should have a reasonable expectation of privacy and if so, which information this would apply to. Therefore, we consider that this proposed measure may, in fact, have positive benefits on users' and other persons' rights to privacy and help to safeguard them.

Impacts on services

18.128 In addition to the costs of measures UR2 (a), (b), (c) and (d) discussed above, providers will incur additional costs for measure UR2 (e), in relation to making an explanation of whether the service provider notifies users when their content or account is complained about, and, if so, what information the notification includes, easily accessible during the reporting process.

18.129 We are not being prescriptive about where this information should be located, but there will be a cost in making it accessible during the complaints process. Providers of larger services may choose to develop an interstitial or banner with this information when a user clicks on the reporting tool. Providers of smaller services may prefer to display or link to the information at the point that the user submits the complaint. We expect associated costs to be largely incurred in design, quality assurance, and testing. We have estimated the direct cost of this measure would take approximately 1 day to 2 weeks of software engineering time, with up to an equivalent amount of non-engineering time. Using our assumptions on labour costs required for this type of work set out in Annex 12, we would expect the one-off direct costs to be somewhere in the region of £400 to £9,000, with annual maintenance costs at 25% of this being around £100 – £2,250 per annum. We would expect that providers of smaller services with simpler complaints process would incur costs to towards the lower end of this range as they can deploy the simpler approach described above in making this information available.

Which providers we propose should implement this measure

18.130 The evidence discussed above suggests that measure UR2 (e) would help providers of U2U services to meet their duties to operate transparent complaints procedures in relation to all types of complaints, where the complainant is a child. While there are some costs of implementing this measure, we consider that they are likely to be proportionate for all providers of services likely to be accessed by children, particularly since they are likely to be significantly lower for providers of smaller services with simpler complaints processes.

18.131 We therefore have provisionally concluded to recommend this measure to all U2U services that are likely to be accessed by children in relation to complaints about suspected illegal content and content considered harmful to children. We are proposing to include this measure in both our draft Illegal Content Codes and our draft Children's Safety Codes.

[506] Article 5 UK GDPR and further guidance on this can be found on the ICO website A guide to the data protection principles. [accessed 19 April 2024].

Other options considered in relation to measure UR2

18.132 To ensure complaints processes are easy to use, we considered whether UK users should be able to complain in languages other than English.[507] However, we decided it was not helpful to be prescriptive about which languages providers should accept complaints in, as appropriate languages are likely to vary depending on the user base and nature of content hosted by individual services. We believe that our proposal that providers should ensure information and processes relating to complaints are comprehensible and accessible, taking into account the needs of their users, should address any need for services to accept complaints in languages other than English, including, for example, in Welsh, where appropriate.

18.133 We considered whether to recommend that service providers collaborate with specialist children's organisations when designing their complaints processes.[508] However, we are aware that there are a limited number of children's organisations across the UK, and to recommend this could be burdensome for them. Given the vast number of services in scope of the Act, we do not think it would be possible for all providers to collaborate with these organisations. That said, we are aware that many providers already seek expert advice when designing their tools to protect children, and we would encourage them to continue to do so where they consider it appropriate. We believe that our proposal that providers should ensure their complaints processes are comprehensible and accessible to users, including children, will secure that complaints processes can be easily used by children, while allowing providers the flexibility to determine how best to achieve that. We will engage directly with children's organisations as part of our stakeholder engagement and encourage them to respond to our consultation proposals.

18.134 We considered whether to recommend standardised categories of content that providers should present to users when complaining.[509] While we acknowledge that setting out different categories of complaints can be useful, and we want to encourage providers to include these where appropriate, we think that providers are best placed to determine which categories of content harmful to children are most appropriate for their particular user base, risks and terms of service. Furthermore, we have not seen consistent evidence about which categories are most helpful for users.[510] However, we welcome further evidence on this and may revisit this measure in future.

18.135 In light of evidence that children sometimes find categories are not appropriate for the types of content they want to report, we also considered whether to recommend that where providers present users with categories of content, they should offer children an 'other' category. However, we have concerns that this could cause unintended consequences for services' backend prioritisation processes, for instance making it harder to identify high priority complaints categorised as 'other' and handle them appropriately. We believe that our proposal recommended above, that providers should ensure information relating to

[507] In their response to our 2023 Protection of Children Call for Evidence, 5Rights suggested services should provide information on moderation and redress in local languages. 5 Rights response to 2023 Protection of Children Call for Evidence.
[508] Suggested by Refuge. Refuge response to 2023 Protection of Children Call for Evidence.
[509] Suggested by Samaritans. Samaritans response to 2023 Protection of Children Call for Evidence.
[510] For example, some participants in our research into children's experiences of cyberbullying felt that categories could be overly restrictive. Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.

complaints processes is comprehensible and accessible to users, including children, would also apply to any categories presented to children when complaining.

18.136 We also considered whether to recommend that providers run information campaigns to raise awareness of complaints processes and encourage users to complain.[511] However, we consider that User Support Measure US6 in Section 21, relating to the information providers should make accessible to children and parents, will improve transparency in a more proportionate, less prescriptive manner.

Provisional conclusion

18.137 Given the harms this measure seeks to mitigate in respect of content harmful to children, as well as service providers duties to operate systems and processes that allow people in the UK to easily report content harmful to children and to operate complaints procedures that are easy to access, easy to use (including by children) and transparent. We consider this measure appropriate and proportionate to recommend for inclusion in the draft Children's Safety Codes. For the draft legal text for this measure, please see PCU C2 and C3 in Annex A7, and PCS C2 in Annex A8.

18.138 We also consider the additional element of this measure appropriate and proportionate to recommend for inclusion in the draft Illegal Content Codes. For the draft legal text for this measure, please see X1 in Annex A9.

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

Explanation of the measure

18.139 Under the Act, providers of U2U and search services have duties to operate complaints processes that provide for appropriate action to be taken by the provider in response to relevant complaints for services likely to be accessed by children. They also have duties to operate transparent complaints procedures.[512]

18.140 As discussed in the 'Interaction with Illegal Harms' section above, as part of this measure we are making a proposal that mirrors an equivalent proposal in the draft Illegal Harms Codes. We are also proposing to add an additional element to this measure to protect children from harm in light of new evidence regarding the barriers children face to reporting. We discuss the equivalent proposal and then the additional element in turn below.

Equivalent proposal

18.141 In our Illegal Harms Consultation, we set out evidence that suggests complainants can often wait a long time to hear the outcome of their complaints and in many cases receive no response at all.[513] This is further supported by research discussed in the 'Reasons why complaints processes are underused' section above, and in Section 7.11, Governance, systems and processes. The evidence suggests that this can create an impression that providers take no action in response to complaints, undermining trust in complaints

[511] See Ygam response to 2023 Protection of Children Call for Evidence.
[512] See sections 21(2)(b) and (c) and 32(2)(b) and (c) of the Act.
[513] Our Illegal Harms Consultation, Volume 4, Section 16, Measure 3.

processes and discouraging complainants (particularly children) from complaining again in future.[514] It also suggests that complainants do not understand what happens following submission of a complaint.[515] We consider that this implies that some providers are not sufficiently transparent about how their complaints processes work and are not always taking appropriate action in response to complaints.

18.142 We made a recommendation to address this in Section 16, Reporting and complaints, of the Illegal Harms Consultation. We are now provisionally proposing to extend that recommendation to apply also to relevant complaints for providers of services likely to be accessed by children.

18.143 **Measure UR3 (a):** We are proposing that providers of all U2U and search services likely to be accessed by children should acknowledge receipt of each relevant complaint and provide the complainant with an indicative timeframe for resolving the complaint.

18.144 We set out our analysis of the evidence, costs and impacts of this measure in the 'Equivalent proposal' section below. We provisionally consider that this measure is proportionate for providers of all U2U and search services likely to be accessed by children. We set out our reasoning for this below.

Additional element

18.145 Since publishing our Illegal Harms Consultation, we have become aware of additional evidence that demonstrates just how significant a barrier lack of trust in complaints processes is for children, and the impact not receiving a satisfactory communication from services in response can have on their likelihood to complain again in future. In light of this evidence, which we discuss below, we are proposing to recommend a measure that was not included in our Illegal Harms Consultation. The evidence for this measure does not relate to specific harms or types of content, but to children's behaviour and attitudes to complaining in general. We consider that this measure will help to make complaints processes more transparent in relation to all types of complaints. We are therefore proposing to include it in both the draft Illegal Content Codes and the draft Children's Safety Codes.

18.146 **Measure UR3 (b):** We are proposing that providers of all U2U and search services should include in their acknowledgement of each complaint an explanation of what actions the provider may take in response to the complaint and whether the complainant should expect to hear the outcome of their complaint. This could be done by linking to another page which contains this information, for example, in the terms of service or a user resource centre. It would not need to be personalised to the complaint or the complainant.

18.147 As we explain below, we are recommending this measure for providers of all U2U and search services likely to be accessed by children in relation to relevant complaints for all services. We are also recommending it for providers of all U2U and search services likely to be accessed by children in relation to additional relevant complaints for services likely to be accessed by children. We discuss the rationale for this, and the evidence and impacts of this measure in the 'Additional element' section below.

[514] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.
[515] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.

Equivalent proposal: Measure UR3 (a)

Measure UR3 (a): acknowledge receipt of each relevant complaint and provide the complainant with an indicative timeframe for resolving the complaint.

Effectiveness at addressing risks to children

18.148 The analysis below does not differ substantially from the discussion in Section 16, Reporting and complaints, of our Illegal Harms Consultation. However, we have added new evidence where appropriate, and updated our language in places to clarify our recommendations.

18.149 If complainants do not feel that their complaints are being dealt with, there is a risk that they may be discouraged from complaining again in future.[516] This means it may take longer for service providers to be made aware of content harmful to children present on their services. Ofcom's 2024 research into children's experiences of violent content online found that doubts about the impact of reports, and lack of feedback or acknowledgement of reports, were barriers discouraging children from reporting violent content.[517]

18.150 Ofcom's 2024 research on reporting behaviours and attitudes in children found that receiving an acknowledgement of their report increases children's confidence in reporting and encourages them to report again in future. Participants in this research also said that responses should be speedy, preferably within 48 hours, stating they felt reports are often ignored and rarely lead to take down. This suggests timely responses are important to instil confidence amongst children in the reporting process.[518]

18.151 This was echoed in responses to our 2023 CFE. Samaritans called for users who complain about self-harm or suicide content to receive an acknowledgement of their complaint and information on what happens next, and any action taken by the service provider in response.[519] The Executive Office: Good Relations and TBUC Strategy, and the Center for Countering Digital Hate also suggested that services should acknowledge complaints.[520]

18.152 Other respondents to both our 2022 and 2023 CFEs suggested it would be helpful to have greater clarity about the process once a complaint has been made, including timings for handling it.[521] 5Rights said timescales should be proportionate to the seriousness of a complaint, which in some instances may require an immediate response.[522] The End Violence Against Women coalition recommended that services should acknowledge complaints within 24 hours and that complaints should be actioned within a specific timeframe set out in the Terms of Service.[523] Refuge said in its Unsocial Spaces Report that complaints about serious offences should be dealt with within 24-48 hours.[524] In their 2023 report, Parentzone called for Ofcom to require services to respond to complaints in a timely and proportionate manner.[525]

---

[516] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.
[517] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.
[518] Ofcom, 2024. Children's Attitudes to Reporting Content Online.
[519] Samaritans response to 2023 Protection of Children Call for Evidence.
[520] The Executive Office NI, Good Relations and TBUC Strategy response to 2023 Protection of Children Call for Evidence; Center for Countering Digital Hate response to 2023 Protection of Children Call for Evidence.
[521] Catherine Knibbs Ltd – trading as Children and Tech response to 2022 Illegal Harms Call for Evidence'.
[522] 5Rights response to 2022 Illegal Harms Call for Evidence.
[523] End Violence Against Women Coalition response to 2023 Protection of Children Call for Evidence.
[524] Refuge, 2021. Unsocial Spaces Report [accessed December 2023].
[525] Parentzone, 2023. Tools – A false hope [accessed November 2023].

18.153 Ofcom's VSP guidance also highlights the importance of setting timeframes for actioning complaints. This can be useful in developing metrics as a way of demonstrating effective procedures for the handling and resolution of complaints.[526] Meanwhile, experiences in other sectors show that a response within two working days increases confidence in complaints handling processes.[527]

18.154 However, as set out in Section 16, Content moderation for U2U services and in Section 17, Search moderation, we are conscious of the risk of perverse outcomes if we were to suggest a one-size-fits-all approach to deadlines for content moderation processes, including those for complaints. It could incentivise service providers' content moderation teams to prioritise speed over accuracy when reviewing complaints or restrict services providers' ability to apply their resources flexibly as new types of harmful content emerge.

18.155 We consider that complainants' concerns about lack of action are likely to be allayed somewhat by having some indicative idea of the timeframe for their complaint to be processed, and that increased trust in the process would make them more likely to use it. There is a risk that an indicative timeframe would be misunderstood by complainants as a binding deadline, leading to adverse outcomes, such as further undermining trust in providers' handling of complaints if complainants believe the timeframe is not being met. However, we consider that service providers would be able to draft the acknowledgement in such a way that it did not lead to false expectations.

18.156 In addition, a recommendation that service providers provide indicative timeframes to complainants would incentivise services to set timeframes which are appropriately swift and to meet them. As set out above, one of the main purposes of the Act is to secure that transparency and accountability are provided in relation to services, and we think that this measure would help to achieve this.

Rights assessment

Freedom of expression and association

18.157 We do not consider that acknowledging receipt of complaints and providing information about timeframes for handling those complaints would, in and of itself, have any adverse impacts on complainants' (which includes both adult and child users) or service providers' rights to freedom of expression or association. Instead, we think that this proposed measure would likely have a positive impact on the rights of users and other complainants by providing transparency and accountability around the complaints process. This has the potential to encourage complaints, resulting in online spaces becoming safer for children and for errors such as, incorrectly categorising content or assessment of a user's age to be rectified.

18.158 To the extent that our proposed recommendations ask service providers to convey information they might not otherwise convey, there is a potential small impact on service providers' rights to freedom of expression. However, we consider this proportionate in the interests of protecting the rights of users and others (including adults and children).

Privacy

18.159 We are of the provisional view that there are unlikely to be any additional impacts on users' and others' rights to privacy beyond those set out in UR1 and UR2. This is because the

[526] Ofcom, 2020. Video-sharing platform guidance.
[527] Legal Ombudsman Best practice complaint handling guide [accessed September 2023].

proposed measure should not require any additional personal data to be retained or processed than is needed to handle complaints under Measures UR1 and UR2 above. This is on the basis that service providers are complying with the requirements of data protection laws and relevant guidance issued by the ICO.[528]

Impacts on services

18.160 Service providers would incur costs from acknowledging complaints and providing indicative timelines (Measure UR3 (a)). We expect that any services that receive more than a small number of complaints would want to automate this acknowledgement (e.g., through an email or pop-up message). We estimated in our Illegal Harms Consultation that this would require 5 to 50 days of software engineering time, with potentially up to the same again in non-engineering time.[529] Using our assumptions on labour costs required for this type of work set out in Annex 12, we would expect the one-off direct costs to be somewhere in the region of £2,000 to £50,000. There would also be some ongoing costs involved in maintaining this measure. Consistent with our standard assumption, we assume that annual maintenance costs are 25% of the initial set-up costs and therefore in the region of £500 to £12,500 per year. We expect that service providers with less complex systems and governance processes are likely to incur costs at the lower end of this range, which is likely to be the case for providers of smaller services.

18.161 Small and low-risk services that do not receive many complaints may choose to have a manual approach to sending acknowledgements and indicative timelines. For these services we expect the cost to be very low.

18.162 We recognise that all service providers who should apply this measure should also apply Measure 3 in the Reporting and complaints section of the Illegal Harms Consultation. Measure UR3(a) in this section closely mirrors that measure, and we would expect providers to incur only small additional costs from extending that measure to relevant complaints for providers of services likely to be accessed by children, on top of relevant complaints for all services, which were covered in the Illegal Harms Consultation.

Which providers we propose should implement this measure

18.163 We consider that this measure is likely to be proportionate for all providers of U2U and search services likely to be accessed by children. Our analysis suggests the costs would be relatively small. Evidence suggests that users are deterred from complaining by lack of clarity about timelines, meaning services may not be made aware of content harmful to children. This suggests that the benefits of applying this measure to large and risky services could be relatively significant. For services that receive very few complaints, the benefits would be small, but the costs would likely also be at the lower end of our estimates, since service providers could retain a manual process for acknowledging complaints.

18.164 Therefore, we propose to recommend that this measure apply to providers of all U2U and search services likely to be accessed by children.

---

[528] Including the ICO's Age Appropriate Design Code, Information Commissioner's Opinion and any relevant guidance such as the ICO's UK GDPR Guidance.

[529] See our Illegal Harms Consultation, Volume 4, Section 16, Measure 4 for full explanation of costs.

Additional element: Measure UR3 (b)

Measure UR3 (b): include in the acknowledgement of each complaint an explanation of what actions the provider may take in response to the complaint and whether the complainant should expect to hear the outcome of their complaint.

Effectiveness

18.165 As mentioned above, since publishing our Illegal Harms Consultation, we have become aware of additional research which strengthens the argument that children are discouraged from complaining again in future if they are not informed of the outcome of their complaints. This is because the lack of response causes children to believe no action has been taken.[530] This implies that the communication children do or do not receive in response to their complaints has the potential to affect their attitude to complaining again in future.

18.166 One potential solution to address this barrier would be to recommend that service providers communicate the outcome of complaints to children. Evidence discussed in Section 7.11, Governance, systems and processes, suggests that this would increase trust in complaints processes and thereby reduce barriers children face to complaining.[531] Many respondents to our 2023 CFE called for service to inform children of action taken in response to their complaints.[532] Girlguiding suggested that this would make young people feel the service had taken their concerns seriously.[533] In response to complaints about content harmful to children, Catch22 suggested services should send personalised responses.[534] NCMEC meanwhile called for services to offer users the ability to track the progress of their complaints.[535]

18.167 Participants in our 2024 research into children's attitudes to reporting said services should update the user on the progress of their report and next steps, including about when they should expect to receive a response. If no action is taken, participants called for an explanation for why the reported content did not violate the rules. Some participants said they felt upset when they reported content they found inappropriate but were told that it did not break the service's rules. While this can be disappointing, the research found that transparency on the outcome of complaints helps to build children's confidence in the reporting process.[536]

18.168 Recommending that service providers tell complainants the outcome of complaints would reassure them that they were considered and would be likely to encourage children to complain again in future. Communicating outcomes could also help to educate children on what content was and was not violative, which over time could help to improve the quality of complaints. It could also reduce children's harmful encounters online, since research suggests that in the absence of responses, some children feel compelled to check if what they reported is still accessible.[537]

---

[530] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.
[531] See 'User reporting and complaints' section of Section 7.11, Governance, systems and processes.
[532] 5Rights response to 2023 Protection of Children Call for Evidence; Common Sense Media response to Ofcom 2023 Call for Evidence; Refuge response to 2023 Protection of Children Call for Evidence.
[533] Girlguiding response to 2023 Protection of Children Call for Evidence.
[534] Catch22 response to 2023 Protection of Children Call for Evidence.
[535] NCMEC response to 2023 Protection of Children Call for Evidence.
[536] Ofcom, 2024. Children's Attitudes to Reporting Content Online.
[537] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.

18.169 We understand that some providers do this currently in some circumstances. For example, in response to our 2022 CFE, Meta told us that following review of a report on Facebook or Instagram, the person making the report will receive a notification informing them of the outcome of their report.[538] We have learnt from our work regulating VSPs that when Twitch acts on content because of a user report, it notifies that user via email.[539] We also learnt that after a report is made, TikTok may update users on the status and progress of their report in their inbox, and the report outcome may also be viewed in the settings under report records.[540]

18.170 However, recommending services communicate the outcome of complaints may go further than is required to build trust in services' complaints handling. There is also a risk that users who were informed no action was taken would be discouraged from complaining again, even if this decision was entirely legitimate on the part of the service provider. Nor do we have sufficient evidence at this stage of the practicalities and costs of implementing such a requirement at scale for each of the types of complaints that services are required to consider. While we welcome further evidence on the topic to inform our future work, at this stage we are therefore not proposing to require providers of services to communicate the specific outcome of complaints to users.

18.171 Instead, we think it would be possible to help dispel the misconception that lack of communication means providers take no action in response to complaints by including information about how complaints are handled in the acknowledgement of complaints. Specifically, we propose provides should include in their acknowledgement of complaints an explanation what actions the provider may take in response to the complaint. This would improve the transparency of services' complaints handling processes and could help reassure children that services review their complaints. This could increase children's trust in complaints procedures, encouraging them to complain about harmful content so that providers can protect other children from it in future.

18.172 In the interests of transparency and to manage expectations, we also consider that users should be informed of whether they should expect to hear the outcome of their complaint. This information would not need to be personalised but could explain the provider's general position on responding to complaints. We think that this would also help to dispel the perception that lack of communication from the service means their complaint was not reviewed or actioned, increasing trust in services' handling of complaints.

18.173 In the absence of clear evidence for what format and wording would be most effective, we consider that services would be best placed to decide the presentation and language used in their acknowledgement of complaints. However, to achieve the aim of improving transparency and increasing users' trust in complaints mechanisms, in line with Measure UR2 (b) above, services would need to ensure the information included in their acknowledgement of complaints was comprehensible and accessible, including to children.

Rights assessment

Freedom of expression and association

18.174 We do not consider that providing information about what actions the service provider might take and whether complainants would be informed of the outcome would, in and of

[538] Meta response to 2022 Illegal Harms Call for Evidence.
[539] Ofcom, 2022. Ofcom's first year of video-sharing platform regulation.
[540] TikTok, Safety & privacy controls [accessed 22 January 2024].

itself have any adverse impacts on complainants' (which includes both adult and child users) rights to freedom of expression or association. We acknowledge that there is potential for a slight risk of interference with services' rights to freedom of expression by proposing recommendations that include details of information that service providers should provide to complainants. However, we have designed this proposed measure with flexibility that allows service providers to decide how they implement it and provide this information to complainants. We are not proposing to recommend specific wording for acknowledgements of complaints.

Privacy

18.175 We are of the view that there are unlikely to be any significant negative impacts on individuals' rights to privacy, beyond those we have set out above in Measures UR1 and UR2, nor should it require the processing of any additional personal data beyond that which is necessary for those measures or for UR3 (a). We think the proposed measure sets out information that is in line with the Act's objectives on transparency in respect of complaints by making clear what will happen following receipt of a complaint.

18.176 We have designed this proposed measure with flexibility that allows service providers to decide how they implement it and provide this information to complainants. We are not proposing to recommend specific wording that services should use to provide this detail.

18.177 We are also not proposing that complainants should be provided with any personal data (or other information that could be protected by this right) of the subject of the complaint. In implementing this proposed measure, service providers should ensure they comply with data protection laws and familiarise themselves with any relevant guidance issued by the ICO.[541]

Impacts on services

18.178 In addition to the costs of Measure UR3 (a) discussed above, providers will incur an additional incremental cost from providing an explanation of what actions the provider may take in response to the complaint and an explanation of whether the complainant should expect to hear the outcome of their complaint. As this information would not need to be personalised to a given complainant or complaint, we expect providers to incur only a small incremental cost of including this information in the acknowledgement of a complaint, in addition to the costs of sending the acknowledgement and providing timeframes.

18.179 There will also be costs for providers of agreeing what actions they may take in response to complaints and getting these signed off through their internal governance processes. We discuss the costs of this in Measures UR4 (U2U) and UR5 (search) below. Providers will also incur a cost from drafting an explanation of these actions for inclusion in the acknowledgement of complaints, however we consider this cost to be negligible.

18.180 We consider that the cost of agreeing whether complainants should expect to hear the outcome of complaints and drafting an explanation of this for inclusion in the acknowledgement would also be negligible.

Which providers we propose should implement this measure

18.181 The evidence discussed above suggests that Measure UR3 (b) would help providers of all U2U and search services likely to be accessed by children meet their duties to operate

[541] See footnote above.

transparent complaints procedures in relation to all types of complaints. While there are some costs of implementing this measure, we consider that they are likely to be negligible in addition to the cost of acknowledging complaints with an indicative timeframe for resolving the complaint.

18.182 We therefore have provisionally concluded to recommend this measure to all U2U services likely to be accessed by children in relation to relevant complaints for all services, as set out in Section 21(4) of the Act and in relation to relevant complaints for services likely to be accessed by children, as set out in Section 21(5) of the Act.

Other options considered for Measure UR3

18.183 We also considered whether to recommend that service providers offer child users the option to opt out of receiving communications relating to a complaint because, in theory, it could be distressing to be reminded of an encounter with harmful content. However, we have not seen evidence that this is a problem for children in practice. In fact, as we discuss, our evidence consistently suggests that children want more communication from services about their complaints rather than less. For these reasons, we decided not to recommend this measure at this stage.

Provisional conclusion

18.184 Given the harms this measure seeks to mitigate in respect of content harmful to children, as well as service providers' duties to operate complaints processes that are transparent and that provide for appropriate action to be taken by the provider in response to complaints about content harmful to children and other types of complaints, we consider this measure appropriate and proportionate to recommend for inclusion in the draft Children's Safety Codes. For the draft legal text for this measure, please see PCU C4 in Annex A7 and PCS C4 in Annex A8.

18.185 We also consider the additional element of this measure appropriate and proportionate to recommend for inclusion in the draft Illegal Content Codes. For the draft legal text for this measure, please see X2 and Y1 in Annex A9.

Measure UR4: Take appropriate action in response to each complaint – U2U

Explanation of the measure

18.186 The Act requires all providers of regulated U2U services likely to be accessed by children to operate processes that provide for appropriate action to be taken in response to complaints about content harmful to children and other types of complaints.[542] The appropriate action that a provider might take will depend on the type of complaint.

18.187 Taking appropriate action in response to complaints is crucial for enabling complaints processes to realise their potential as a measure to protect children from harm, since it is only when providers react to complaints appropriately that their awareness of harmful content can translate into greater protections for children. It is also very important that

[542] See sections 21(2)(b) of the Act.

providers take appropriate action in response to appeals about wrongful removal or restriction of content or accounts, as this helps to ensure users are treated fairly and their right to freedom of expression is respected. Unless providers take appropriate action in response to complaints by users who are unable to access content because of an incorrect assessment of their age, then users may be unfairly prevented from accessing content through no fault of their own.

18.188 We have therefore considered what 'appropriate action' might mean for providers of U2U services likely to be accessed by children in the context of the different types of complaints envisaged by the Act:

a) Complaints about content harmful to children;
b) Complaints about wrongful restriction/removal of content or account (appeals);
c) Complaints about inability to access content because of an incorrect assessment of the user's age; and
d) Complaints about non-compliance with safety duties protecting children.

18.189 We propose to recommend that providers of all U2U services likely to be accessed by children should take appropriate action in response to these types of complaints. We set out below what we consider this means in practice. For reasons of proportionality, for some types of complaints we propose to recommend different measures for large services or services that are multi-risk for content harmful to children, compared to smaller, low-risk or single-risk services.

Measure UR4 (a): complaints about content harmful to children

18.190 When a provider receives a complaint about content considered harmful to children:

a) if the provider has established a process for content prioritisation and applicable performance targets, it should handle the complaint in accordance with them; or
b) if the provider has no process for content prioritisation and applicable performance targets, it should consider the complaint promptly; and
c) in either case, it should comply with Content Moderation Measure CM1 in Section 16 and Recommender Systems Measures RS1 and 2 in Section 20 regarding the handling of such content.

Measure UR4 (b): appeals

18.191 Measure UR4 (b) (i): When a provider of a large service or a service that is multi-risk for content harmful to children receives an appeal:

a) The provider should have regard to the following matters in determining what priority to give to review of the complaint:

i) the seriousness of the action taken against the user and/or the content as a result of the decision that the content was content harmful to children;[543]

[543] This is a slight change to the wording which we consulted on for the equivalent measure in our Illegal Harms Consultation: there we said only, "the severity of the action taken against the user as a result (etc)". We consider that the addition of a reference to the content resolves a possible ambiguity over whether action taken against a user includes action taken against that user's content. Subject to consultation responses, we propose to include this revised wording in the equivalent measure in our Illegal Harms Statement.

ii) whether the decision that the content was content harmful to children was made by content identification technology;[544]

iii) information that we have recommended the provider collect about the likelihood of false positives generated by the specific content identification technology used, [545] and any other information available about the accuracy of the content identification technology at identifying similar types of content harmful to children;[546] and

iv) the service's past error rate in making judgements about similar kinds of content harmful to children.

b) the provider should set and monitor performance targets relating to the time it takes to determine the appeal and the accuracy of decision making and should resource itself so as to be able to meet those targets; and

c) if, on review, a provider reverses a decision that content was content harmful to children, the provider should:

i) reverse the action taken against the user or in relation to the content (or both) as a result of that decision (so far as appropriate for the purpose of restoring the position to what it would have been had the decision not been made);[547]

ii) where necessary to avoid similar errors in future, adjust the relevant content moderation policies; and

iii) where applicable, and necessary to avoid similar errors in future, take such steps as are within its power to secure that the use of automated content moderation

[544] This is a slight change to the wording which we consulted on for the equivalent measure in our Illegal Harms Consultation: there we referred to "proactive technology" and not "content identification technology". We consider "content identification technology" better reflects the policy objective because such technology when used in response to a user complaint would not be "proactive technology" as defined in the Act, but the harm to the user concerned would be the same. Subject to consultation responses, we propose to include this revised wording in the equivalent measure in our Illegal Harms Statement.

[545] We are not proposing at this time to include any measures in the draft Children's Safety Codes that recommend providers should collect false positive rates for their content identification technologies. However, we may do so in future. We have therefore drafted this measure in such a way that, if we do recommend that in future, we would not need to change this measure to recommend that providers should consider those false positive rates when prioritising appeals.

[546] This is a slight change to the wording which we consulted on for the equivalent measure in our Illegal Harms Consultation: there we said, "and the likelihood of false positives generated by the specific proactive technology used" instead of the wording set out here. This is because while we still think it is appropriate for providers to consider false positive rates where we have specifically recommended they should collect them, we are aware that some services who should apply this measure may not be collecting those metrics and may instead be collecting different information about the accuracy of their content identification technologies. Subject to consultation responses, we propose to include this revised wording in the equivalent measure in our Illegal Harms Statement.

[547] This is a slight change to the wording which we consulted on for the equivalent measure in our Illegal Harms Consultation: there we said, "to the position they would have been in had the content not been judged to be illegal content" instead of the wording used here. This is because we have become aware through responses to our Illegal Harms Consultation that the previous wording had the potential to be misunderstood by providers. We consider the wording used here better reflects the policy objective, which is that the provider should reverse any action taken against the content or user, rather than necessarily restore it to the exact same position it would have been in. Subject to consultation responses, we propose to include this revised wording in the equivalent measure in our Illegal Harms Statement.

technology does not cause the same piece of content to be taken down, down ranked, or restricted again.[548]

18.192 Measure UR4 (b) (ii): When a provider of a service that is neither large nor multi-risk for content harmful to children receives an appeal:

a) the provider should handle it promptly; and
b) if, on review, a provider reverses a decision that content was content harmful to children the provider should:

i) reverse the action taken against the user or in relation to the content (or both) as a result of that decision (so far as appropriate for the purpose of restoring the position to what it would have been had the decision not been made); [549]
ii) where applicable, and necessary to avoid similar errors in future, adjust the relevant content moderation policies; and
iii) where applicable, and necessary to avoid similar errors in future, take such steps as are within its power to secure that the use of automated content moderation technology does not cause the same piece of content to be taken down, down ranked or restricted again.[550]

Measure UR4 (c): complaints about incorrect assessment of a user's age

18.193 Measure UR4 (c) (i): When a provider of a service that we recommend should implement any of the Age Assurance Measures AA3-6 in Section 15 receives a complaint about incorrect assessment of age:

a) the provider should have regard to the following matters in determining what priority to give to the review of the complaint:

i) the seriousness of the restriction applied to the user's account as a result of the assessment of their age;
ii) whether the decision to restrict access to content on the basis of the assessment of their age was made by an age assurance method without human oversight and the likelihood of incorrect assessment by the specific technology used;
iii) the provider's past error rate in making assessments of age of the type concerned; and
iv) any representations made by the complainant as part of the complaint as to the effect of the decision on their livelihood (for example, an adult performer may be unable to their access earnings on an adult website if incorrectly assessed to be under-18).

b) the provider should set and monitor performance targets relating to the time it takes to determine the complaint and the accuracy of decision making and should resource itself so as to be able to meet those targets; and
c) if, on review, a provider reverses a decision to restrict a user's access to content on the basis of an incorrect assessment of their age, the provider should:

[548] This is a slight change to the wording which we consulted on for the equivalent measure in our Illegal Harms Consultation: there we said, "same content" and not "same piece of content". We have changed this to clarify the intention behind this measure. Subject to consultation responses, we propose to include this revised wording in the equivalent measure in our Illegal Harms Statement.
[549] See footnote 547 above.
[550] See footnote 548 above.

i) restore the user's ability to access content on the service to an equivalent position to the one it would have been in had the assessment of age been correct; and

ii) monitor trends in complaints about incorrect assessments of age and use this information to help ensure their age assurance method fulfils the criteria for highly effective age assurance set out in Annex 10 (draft HEAA guidance).

18.194 Measure UR4 (c) (ii): When a provider of a service that we do not recommend should implement any of the Age Assurance Measures AA3-6 in Section 15 receives a complaint about incorrect assessment of age:

a) the provider should handle it promptly; and
b) if, on review, a provider reverses a decision to restrict a user's access to content on the basis of an incorrect assessment of their age, in principle the service should restore the user's ability to access content on the service to an equivalent position to the one it would have been in had the assessment of age been correct.

Measure UR4 (d): complaints about non-compliance with the safety duties protecting children

18.195 For these types of complaints we consider:

a) the provider should establish a triage process for relevant complaints about non-compliance with the safety duties protecting children with a view to protecting users from harm. A responsible person, team or function should be nominated to lead this triage process and ensure relevant complaints for services likely to be accessed by children reach the most relevant function or team.
b) relevant complaints for services likely to be accessed by children should be dealt with:

i) in a way that protects users' and the provider's compliance with other applicable laws in question;
ii) within timeframes the provider has determined are appropriate; and
iii) in accordance with the other appropriate action recommendations set out above.

18.196 These recommendations largely codify the requirement in the Act. Where we make different recommendations for providers of services of different sizes or risk profiles, this is because we consider the minimum level of appropriate action is different depending on the type and level of risk of a service. This is because of the different volumes and types of complaints such services are likely to receive.

18.197 Other than a few minor changes to the wording we are suggesting above, these proposals mirror equivalent ones in the draft Illegal Content Codes, which recommend all providers of U2U services take appropriate action in response to other types of complaints, such as complaints about suspected illegal content. Providers applying both proposed measures may operate a single complaints process to provide for appropriate action in response to various different types of complaints, should they wish to do so.

Justification for the measures

18.198 For the most part, the analysis below does not differ substantially from the discussion in Section 16, Reporting and complaints, of our Illegal Harms consultation. However, we have updated our language in places to clarify our recommendations and make our rationale easier to follow. In some places we have also changed our reasoning to make it pertinent to relevant complaints for services likely to be accessed by children, as set out in section 21(5) of the Act.

Measure UR4 (a): complaints about content considered harmful to children

18.199 Once a complaint about content considered harmful to children has been received, it should enter the provider's content moderation function. As set out in Section 16, this means that all providers will need to handle the complaint in accordance with Content Moderation Measure CM1 as well as Recommender Systems Measures RS1 and 2 in Section 20.

18.200 The Act sets out categories of content harmful to children, which providers have duties to use proportionate systems and processes designed to protect children from (PC and NDC) or prevent them from encountering (PPC).[551] Providers have a choice about how they assess content to meet these duties. If providers use different categories of content in their terms of service from those used in the Act, but they are nonetheless confident that their alternative categorisation secures the same protections for children as required by the Act, then they may assess complained about content using the categories in their terms of service. If the categories in their terms of service do not secure the protections for children required by the Act, then the provider will need to assess complained about content using the categories defined in the Act and explained in Volume 3, Section 8, Ofcom's Guidance on Content Harmful to Children.

18.201 Providers of large services (regardless of their risk), and services that are multi-risk for content harmful to children, would also need to handle the complaint in accordance with their prioritisation process and performance targets under U2U Content Moderation Measures CM3 and 4 in Section 16.

18.202 Providers of smaller services that are not multi-risk for content harmful to children may not receive many, if any, complaints across diverse types of potentially harmful content, and may therefore not require prioritisation processes and performance targets in order to deal with complaints effectively. We are therefore not recommending in the U2U Content Moderation section that these service providers must establish prioritisation processes and performance targets. We consider that if a provider of a smaller service that is not multi-risk has chosen to establish a prioritisation process and performance targets, it would be appropriate to abide by them. But a provider which has none would nevertheless be expected to process all complaints received promptly.

18.203 If the provider determines that the content was either content harmful to children or otherwise captured under a relevant term of service, the provider should then take steps to provide for compliance with the safety duties protecting children.

Measure UR4 (b): appeals

18.204 The Act requires providers to enable users to complain if their content is taken down or restricted for being content harmful to children (including due to the use of proactive technology), or their account receives a warning, suspension or ban as a result of posting content harmful to children. As defined above, we refer to these types of complaints as 'appeals'.

18.205 Some providers may choose to run appeals through their main content moderation function. Others may establish a separate team. In either case, there are questions about how appeals should be prioritised and how quickly they should be handled.

18.206 Providers of large services and services that are multi-risk for content harmful to children may receive a large volume of appeals, possibly across different types of potentially harmful

[551] Section 12(3) of the Act.

content. As a result, there is a risk that users may be harmed if they do not consider appropriate prioritisation in advance. We therefore provisionally consider that providers of large services and services that are multi-risk for content harmful to children should have regard to the matters set out under Measure UR4 (b) (i) in the 'Explanation of the measure' section above when determining what priority to give an appeal.

18.207 For providers of services that are not large and are not multi-risk, we provisionally think that there is no need to make detailed recommendations in Codes on prioritisation. We set out our reasons for this in Section 16, Content moderation for U2U services.

18.208 On the timeliness of considering appeals, for all the reasons set out in Section16, Content moderation for U2U services, we do not consider it appropriate for Ofcom to make specific recommendations. For providers of services which are not large and are not multi-risk, which we expect will not receive many appeals, we consider it will be sufficient to say that appeals should be determined promptly.

18.209 However, we consider that taking this approach for providers of large services and services that are multi-risk for content harmful to children could create perverse incentives and lead to user harm, for example by incentivising providers to resolve complaints quickly rather than accurately. We therefore propose to recommend that such providers should set and monitor targets for speed and accuracy for the determination of appeals.[552] Our reasoning for this is the same as set out in relation to content moderation decisions in Section 16, U2U Content Moderation. Similar recommendations in Section 16 as to monitoring and resourcing would apply in relation to these too, for the reasoning given there.

18.210 We consider that if, on review, a provider reverses a decision that content was content harmful to children, the provider should:

a) reverse the action taken against the user or in relation to the content (or both) as a result of that decision (so far as appropriate for the purpose of restoring the position to what it would have been had the decision not been made); and
b) where necessary to avoid the same error occurring again in future, adjust the relevant content moderation policies.

18.211 The policy intention behind point (a) here is that the provider should reverse the action they took against the user or their content, for example by removing any restriction placed on it or reinstating it if it had been removed from the service. We recognise that it may not be practical to restore the content to the exact position it would have been in had it not been incorrectly judged to be content harmful to children (e.g., the same position in a recommender feed) and this is not the intention of this recommendation.

18.212 It is possible that automated content identification technology may be involved in a restriction or removal decision. We therefore propose that if on review, a provider reverses a decision that content was content harmful to children, then where necessary to avoid the same error occurring again in future, the provider should take such steps as are within its power to secure that the use of automated identification moderation technology does not cause the same piece of content to be taken down, down ranked or restricted again.

[552] There are a number of different ways providers could monitor performance against accuracy targets. For example, providers might want to select a sample of appeal decisions for a second review and track the number of decisions that were overturned by a second reviewer. We do not propose to be prescriptive about how providers should monitor performance against accuracy targets.

Measure UR4 (c): complaints about incorrect assessment of a user's age

18.213 The Act requires all providers of U2U services likely to be accessed by children to enable users to complain if they are unable to access content because of an incorrect assessment of their age.

18.214 For providers of services we recommend should implement any of our Age Assurance Measures AA3-6, we deem that correct age assessments are important for those measures to have the intended effect, protecting children whilst still allowing adults to access lawful content. We recommend in Section 15 that those providers implement age assurance that is highly effective, which reduces the likelihood of incorrect age assessments.[553] However, some likelihood of incorrect assessments would inevitably still exist, and it can adversely affect the rights of users to access content that they should be able to access. It is therefore important that services who implement highly effective age assurance in line with our recommendations should take the additional steps set out under Measure UR4 (c) (i) in the 'Explanation of the measure' section above.

18.215 On the timeliness of considering such complaints, for all the reasons set out in Section 16, Content moderation for U2U services, we do not consider it appropriate for Ofcom to make specific recommendations. Instead, we propose to recommend that providers who we recommend implement highly effective age assurance should set and monitor performance targets relating to the time it takes to determine the complaint and the accuracy of decision making and should resource themselves so as to be able to meet those targets as proposed in this measure. [554] These steps aim to ensure that complaints about age assessments are prioritised appropriately and resolved swiftly, to a high degree of accuracy.

18.216 For providers of services that we are not recommending implement highly effective age assurance at this time, we provisionally think that there is no need to make detailed recommendations in Codes. This is because while we still consider it important that they resolve any age assessment-related complaints swiftly, we do not consider it proportionate to recommend the extra steps related to prioritisation processes and performance targets. For these providers we therefore consider that complaints should be determined promptly and in accordance with data protection law.[555]

18.217 We consider that if, on review, a provider reverses a decision to restrict a user's access to content on the basis of an incorrect assessment of their age, the provider should restore the user's ability to access content on the service to an equivalent position to the one it would have been in had the assessment of age been correct.

18.218 As set out in Annex 10 (draft HEAA guidance), the effectiveness of a method of age assurance depends on how it is implemented. To help mitigate the risk of systemic or repeated errors in the assessment of users' ages, if we recommend the provider of service should implement any of Age Assurance Measures AA3-6, we consider the provider should monitor trends in complaints about incorrect assessments of age and use this information to

[553] Fulfilling the criteria of technical accuracy, robustness, reliability and fairness should reduce the likelihood of inaccurate age assessments. See Annex 10, Draft HEAA guidance for more information.
[554] There are a number of different ways providers could monitor performance against accuracy targets. For example, providers might want to select a sample of complaints about incorrect assessment of a user's age for a second review and track the number of decisions that were overturned by a second reviewer. We do not propose to be prescriptive about how providers should monitor performance against accuracy targets.
[555] Further guidance on this can be found on the ICO website - GDPR Guidance and Resources - Individual rights.

help ensure their age assurance method fulfils the criteria for highly effective age assurance set out in Annex 10. Providers should familiarise themselves with data protection legislation and how to apply it to their age assurance method, for example by taking legal advice when needed and consulting ICO guidance.

18.219 We consider that if, on review, a provider reverses a decision to restrict a user's access to content on the basis of an incorrect assessment of their age, in principle the provider should restore the user's account to an equivalent position to the one they would have been in had the assessment of age been correct.

Measure UR4 (d): complaints about non-compliance with the safety duties protecting children

18.220 The Act also requires services to enable users and affected persons to complain if they consider the provider is not complying with a safety duty protecting children. We note that there is a significant risk of overlap between complaints about compliance with the safety duties protecting children and complaints about content harmful to children, or about wrongful restriction/removal of content or account. Where a complaint falls into one of those categories as well as this, we provisionally consider it appropriate for the provider to handle it in accordance with our proposed recommendations for those complaint types. However, we do not think we need to specify this in a specific measure, since this is already captured by the measures discussed above.

18.221 We provisionally think the appropriate action for providers in relation to complaints concerning compliance with safety duties protecting children would be to establish a triage process aimed at protecting users and affected persons from harm, including harm to their rights, such as to freedom of expression and privacy (see Rights assessment below). A responsible person, team or function for such complaints should be nominated to lead this triage process and ensure complaints reach the most relevant function or team. They should be dealt with in a way that protects users and the provider's compliance with other applicable laws in question, within timeframes the provider has determined are appropriate, and in accordance with our other proposed Code measures relating to complaints.

18.222 At this stage, we are not in a position to predict with sufficient certainty the many different types of complaint that may be submitted to services relating to compliance with the safety duties protecting children, or to set out what action would be appropriate in response to them. Consequently, we are not currently proposing to make detailed recommendations about what final action may be appropriate for these complaints, although we will keep this position under review.

Rights assessment

18.223 This proposed measure recommends that providers of all U2U services likely to be accessed by children, take appropriate action in response to each type of complaint.

18.224 The duty for service providers to take appropriate action in response to complaints is a requirement of the Act, and not of this proposed measure, and we are giving providers flexibility as to precisely how they implement this and what action they take.

Freedom of expression and association

18.225 We do not consider that service providers taking appropriate action in response to complaints would have adverse impacts on complainants' (which may include both adults and children) rights to freedom of expression or association.

18.226 Instead, our view is that this proposed measure would likely have a positive impact on the right to freedom of expression as it is aimed at providers reviewing decisions about content harmful to children in a manner that reflects its own policies on prioritisation and performance targets, or to consider complaints promptly where a service does not have these policies and targets in place. We consider this recommendation provides reassurance to complainants that incorrect decisions can be rectified, reinforcing the Act's objectives to protect children from this content and with the result that fewer children would likely be exposed to content harmful to them. The benefits to children would be that online spaces are made safer for children by reducing the likelihood and period that content harmful to children is present on the service, positively impacting children's rights to freedom of expression and freedom of association as children would be able to engage more safely with communities and content online.

18.227 Our proposal supports the premise that users and others should not be subjected to any detriment if their appeal is upheld, recommending that content or accounts are reinstated to the position they would have been if the content had not been incorrectly categorised as content harmful to children and steps are taken to ensure similar errors are not made in future. It may also mitigate any impact on the user's right to freedom of expression or association where the service provider overturns their previous (incorrect) decision on appeal, giving the user a mechanism for redress.

18.228 We think that this proposed measure would likely have a positive impact on the right to freedom of expression as service providers reviewing decisions made to assess the age of a user with the result that their access to the service or content is restricted, can rectify errors made and ensure that users are granted access to services and content in line with their rights to freedom of expression and freedom of association. The complaints process may also mitigate the impact on the adult users' rights to freedom of expression and freedom of association by giving the user a mechanism for redress and providing a route to rectify negative impacts by allowing adult users access to the service.

18.229 We do not think there would be a negative impact on service providers' rights to freedom of expression by proposing recommendations that set out what we think appropriate action will entail. We have designed this proposed measure with flexibility that allows service providers to decide how they implement it and what the outcomes of complaints would be, provided outcomes are in line with the Act's objectives to protect children from content harmful to them.[556] We think that there would be a positive impact on service providers' rights to freedom of expression by recommending they set out clear processes and providing them with the flexibility to determine their own boundaries, so long as they comply with the requirements of the children's safety duties set out in the Act.

Privacy

18.230 We think that our proposals could have a positive impact on the right to privacy by providing greater transparency and accountability around decisions that are made in relation to content harmful to children, non-compliance with the safety duties protecting children or incorrect assessments of age. It would, in our provisional view, incentivise service providers to ensure more decisions are taken correctly due to the resources and time required to consider complaints. It should enable individuals to have some reassurance of the steps that the service provider will take and what information is needed from them to investigate the

[556] Section 12 of the Act.

complaint. In implementing this proposed measure, service providers should ensure they comply with data protection laws and familiarise themselves with any relevant guidance issued by the ICO.[557] It would also provide a clear route for individuals to challenge decisions that may result in service providers processing inaccurate personal data about them.[558]

18.231 In processing users' personal data for the purposes of this measure, including any additional information required to take appropriate action in relation to the complaint, we consider that service providers can and should implement the measure in a way which minimises the amount of personal data that is processed, in line with the principle of data minimisation.[559]

18.232 We therefore consider that the impact of the proposed measure on individuals' (including adults' and children's) rights to privacy to be relatively limited, and potentially overall positive. It is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

Impacts on services

18.233 The costs of taking appropriate action for complaints will vary across different types and sizes of services, and for services with different levels of risk. While we expect the costs could be very significant for some service providers, we believe they derive, in large part, from duties in the Act.

18.234 We are proposing to mitigate the risk of imposing unnecessary costs of our recommendations by allowing service providers flexibility to set their own timescales for resolving complaints. This will help to ensure that the costs incurred are proportionate to the nature and risk profile of the service.

18.235 It should also be noted that if complaints about content harmful to children are routed through a service's content moderation function, the costs of taking appropriate action in that case could be regarded as part of content moderation.

18.236 Furthermore, while we recognise that it depends in part on the nature of the service, we would generally expect the volume of complaints a service provider receives to increase with the size of the service, the risks on the service, the volume of content being shared by users, and the number of content moderation and age assessment decisions being taken by the provider. This means that the highest costs will be incurred by the providers of the largest services who are most likely to be able to absorb them, and who we expect would see the greatest benefits from implementing these recommendations.

18.237 The costs of establishing and running a triage process to ensure that relevant complaints about non-compliance with the safety duties protecting children reach the most relevant function will also depend on the nature of the service, but we expect these costs to scale with the size and complexity of a service, and the volume of complaints.

18.238 Services whose current complaints policies do not meet the measure will also incur costs when adapting their policies regarding what action they will take in response to complaints,

---

[557] As suggested in Measure UR1.

[558] Individuals have the right to have this inaccurate personal data rectified under data protection laws and the ICO have issued guidance on this Right to rectification. [accessed 19 April 2024].

[559] The ICO has published Guidance on the data protection principles. A guide to the data protection principles. [accessed 19 April 2024].

and ensuring they are compliant. Services with more complex governance processes are likely to incur greater costs when agreeing these policies.

18.239 To implement Measure UR4 (b), large and multi-risk services would need to develop prioritisation frameworks and set and monitor performance targets for appeals. We believe that there would be similar activities and therefore costs involved as those described in the context of Content Moderation Measures CM3 and CM4. We consider that there are likely to be some overlaps in the processes required for CM3/CM4 and measure UR4 (b) proposed here, which may imply some cost savings for services.

18.240 We have also considered impacts on service providers who are recommended to apply any of Age Assurance Measure AA3-6 and therefore have some additional recommended steps under Measure UR4 (c) (i). There will be costs to prioritise complaints, set and monitor performance targets, and monitor trends. However, these steps should also have countervailing benefits for services, contributing to user satisfaction (among users who make complaints) and helping services continue to ensure their approach to age assurance is highly effective, as per Measure UR4 (c) (i) part (c) in the 'Explanation of the measure' section above. We expect that the cost of implementing this measure could be lower than for similar measures for complaints or appeals that relate to content, which may be more diverse in nature (e.g. complaints/appeals covering many different kinds of harmful content, different media, different ways in which content is encountered on the service) and therefore may require more sophisticated prioritisation and performance tracking processes.

18.241 We recognise that all service providers who should implement Measure UR4 should also implement the related proposed measure in our Illegal Harms consultation.[560] We consider that there will be scope for significant cost savings where providers use the same systems and processes to provide for appropriate action in relation to the types of complaints covered by the draft Illegal Content Codes and the draft Children's Safety Codes.

18.242 Additionally, we have considered the potential added complexity for all kinds of services in making judgements about content harmful to children. However, as set out above the Act does not necessarily require service providers to make judgements about content harmful to children if they are satisfied that their terms of service or community guidelines deal with content that would be considered content harmful to children under the Act. To the extent that new judgements about content harmful to children are required, this is down to the requirements of the Act.

Which providers we propose should implement this measure

18.243 Due to the fact that the reporting and complaints duties apply to providers of all in-scope services likely to be accessed by children, we have proposed setting out broad features (as opposed to specific ones) that we recommend providers consider when designing their reporting and complaints processes. We believe we have approached this in a way that seeks as far as possible to elucidate the basic legal requirements set out in the Act.

18.244 On this basis, we believe our proposals regarding complaints about content harmful to children (Measure UR4 (a)) and about non-compliance with the safety duties protecting

[560] Our Illegal Harms Consultation, Volume 4, Section 16, Measure 4.

children (Measure UR4 (d)) are proportionate and suitable for providers of all U2U services likely to be accessed by children.

18.245 For appeals Measure UR4 (b), while we propose that this would apply to providers of all U2U services likely to be accessed by children, we propose to make different recommendations for providers of large services and services that are multi-risk for content harmful to children compared to providers of other services. This is because providers of services that are large or multi-risk for content harmful to children are likely to receive a high volume of these types of complaints across a range of different types of content that may be harmful to children. We consider that unless providers of these services consider prioritisation and performance targets in advance, there is a risk that they will be unable to take appropriate action in response to large volumes of complaints. We consider that the benefits of adopting a prioritisation framework and setting and monitoring performance targets in these cases are sufficiently important for them to incur the costs of doing so, in order to be able to take appropriate action in response to appeals given the larger volume of these types of complaints these providers are likely to receive.

18.246 Providers of services that are smaller and are not multi-risk are generally less likely to receive large volumes of complaints across a diverse set of content types that may be harmful to children. As such we do not think it is necessary to recommend that they should establish prioritisation processes or set and monitor performance targets. Given there are likely to be costs involved in implementing these, we do not believe that the potential benefits are large enough to justify these costs to such services. Rather, we think it is sufficient and proportionate to recommend that providers of these services should handle these complaints promptly. We consider this is the minimum necessary for providers of these services to comply with their duty under the Act.

18.247 For complaints about incorrect assessment of age (Measure UR4 (c)), while we propose that this would apply to providers of all U2U services likely to be accessed by children, we propose to make different recommendations for providers who should apply any of Age Assurance Measures AA3-6 in Section 15. Again, while there are likely to be costs incurred as a result of these recommendations, we consider that the benefits of adopting a prioritisation framework and setting and monitoring performance targets and trends in these cases are sufficiently important for them to incur the costs of doing so. These steps aim to ensure that complaints about age assessments are prioritised appropriately and resolved swiftly, to a high degree of accuracy to protect the rights of users to access content that they should be able to access.

18.248 For providers of services which need not apply of any of Age Assurance Measures AA3-6 we do not consider that the benefits are large enough to justify the costs of these additional steps to these services and consider that it is sufficient and proportionate to recommend that providers of these services should handle these complaints promptly and in accordance with data protection law.

18.249 See Section 16, Content Moderation for U2U services, for further discussion of this approach.

Other options considered

18.250 Some stakeholders called for us to be prescriptive about actions a service should take in response to complaints, for instance temporarily suspending access to all reported content

while a review is carried out.[561] We consider that given the wide range of services in scope, it would not be appropriate to be prescriptive about how services should handle complaints and instead propose to allow services the flexibility to handle complaints in line with their policies and (where applicable) KPIs, which can be tailored to their size, userbase and the nature of the content they host. With regards to suspending reported content pending review specifically, we propose under Recommender Systems Measures RS1 and RS2 in Section 20, that to protect children, recommender systems should be instructed not to recommend content that is likely to be PPC to children and to downrank content that is likely to be PC (and potentially, subject to the outcome of the consultation, NDC), whether it has yet been confirmed as such or not through the content moderation process. We consider that these proposed measures would help ensure that such content is not promoted to children while the content moderation process takes place.

Provisional conclusion

18.251 Given the harms this measure seeks to mitigate in respect of content harmful to children, as well as service providers' duties to operate processes that provide for appropriate action to be taken in response to complaints about content harmful to children and other types of complaints, we consider this measure appropriate and proportionate to recommend for inclusion in the draft Children's Safety Codes. For the draft legal text for this measure, please see PCU C5-C11 at Annex A7.

Measure UR5: Take appropriate action in response to each complaint – Search

Explanation of the measure

18.252 The Act requires all providers of regulated search services likely to be accessed by children to operate processes that provide for appropriate action to be taken in response to complaints about content harmful to children and other types of complaints. The appropriate action that a provider might take will depend on the type of complaint.

18.253 Taking appropriate action in response to complaints is crucial for enabling complaints processes to realise their potential as a measure to protect children from harm, since it is only when providers react to complaints appropriately that their awareness of harmful content can translate into greater protections for children. It is also very important that providers take appropriate action in response to appeals about content being wrongfully downranked or no longer appearing in search results, as this helps to ensure website owners are treated fairly and their right to freedom of expression is respected. Unless providers take appropriate action in response to complaints by users who are unable to access content because of an incorrect assessment of their age, then users may be unfairly prevented from accessing content through no fault of their own.

18.254 We have therefore considered what "appropriate action" might mean for providers of search services likely to be accessed by children in the context of the different types of complaints envisaged by the Act:

[561] The Executive Office NI, Good Relations and TBUC Strategy response to 2023 Protection of Children Call for Evidence. Your Data Key response to 2023 Protection of Children Call for Evidence.

a) Complaints about search content that is harmful to children;
b) Complaints by website owners about measures taken that result in their content being wrongfully downranked or no longer appearing in search results (for example, as a result of filtering, deindexing or delisting) (appeals);
c) Complaints about incorrect assessment of the user's age; and
d) Complaints about non-compliance with safety duties protecting children.

18.255 We propose to recommend that providers of all search services likely to be accessed by children should take appropriate action in response to these types of complaints. We set out below what we consider this means in practice.

Measure UR5 (a): complaints about content harmful to children

18.256 When a provider receives a complaint about content considered harmful to children:

a) if the provider has established a process for search moderation prioritisation and applicable performance targets, it should handle the complaint in accordance with them; or
b) if the provider has no process for search moderation prioritisation and applicable performance targets it should consider the complaint promptly; and
c) in either case, it should comply with Search Moderation Measures SM1 and 2 in Section 17 regarding handling of such content.

Measure UR5 (b): appeals

18.257 Measure UR5 (b) (i): When a provider of a large general search service or a search service that is multi-risk for content harmful to children (including multi-risk vertical search services) receives an appeal:

a) The provider should have regard to the following matters in determining what priority to give to the review of the complaint:
   i) the seriousness of the action taken against the website owner as a result of the decision that the content was content harmful to children;[562]
   ii) whether the decision that the content was content harmful to children was made by content identification technology;[563]
   iii) information that we have recommended the provider collect about the likelihood of false positives generated by the specific content identification technology used,[564] and any other information available about the accuracy of the content identification technology at identifying similar types of content harmful to children;[565] and
   iv) the provider's past error rate in making judgements about similar kinds of content harmful to children of the type concerned.
b) the provider should set and monitor performance targets relating to the time it takes to determine the appeal and the accuracy of decision making and should resource itself so as to be able to meet those targets; and
c) if, on review, a provider reverses a decision that content was content harmful to children, the provider should:

[562] See footnote 543 above.
[563] See footnote 544 above.
[564] See footnote 545 above.
[565] See footnote 546 above.

v) reverse the action taken against the website owner or in relation to the content (or both) as a result of that decision (so far as appropriate for the purpose of restoring the position to what it would have been had the decision not been made);[566]

vi) where necessary to avoid similar errors in future, adjust the relevant content moderation policies; and

vii) where necessary to avoid similar errors in future, take such steps as are within its power to secure that the use of automated content moderation technology does not cause the same piece of content to be downranked or removed from search results (through whatever technical means) again.[567]

18.258 Measure UR5 (b) (ii): When a provider of a smaller, low-risk or single-risk general search service or large, low-risk or single-risk vertical search service receives an appeal:

a) the provider should handle it promptly; and

b) if, on review, a provider reverses a decision that content was content harmful to children, the service should:

i) reverse the action taken against the website owner or in relation to the content (or both) as a result of that decision (so far as appropriate for the purpose of restoring the position to what it would have been had the decision not been made)[568]

ii) where necessary to avoid similar errors in future, adjust the relevant content moderation policies; and

iii) where necessary to avoid similar errors in future, take such steps as are within its power to secure that the use of automated content moderation technology does not cause the same piece of content to be downranked or removed from search results (through whatever technical means) again.[569]

Measure UR5 (c): complaints about incorrect assessment of a user's age[570]

18.259 When a provider receives a complaint about inability to access content because of incorrect assessment of a user's age:

a) the provider should handle it promptly; and

b) if, on review, a provider reverses a decision to restrict a user's access to content on the basis of an incorrect assessment of their age, the provider should restore the user's ability to access content on the service to an equivalent position to the one it would have been in had the assessment of age been correct.

Measure UR5 (d): complaints about non-compliance with the safety duties protecting children

18.260 For these types of complaints we consider:

a) the provider should establish a triage process for relevant complaints about non-compliance with the safety duties protecting children with a view to protecting users

[566] See footnote 547 above.
[567] See footnote 548 above.
[568] See footnote 547 above.
[569] See footnote 548 above.
[570] We are not at this time proposing to recommend that providers of search services should operate age assurance to meet their safety duties protecting children (see Section 15). However, providers may nonetheless choose to do so, and the Act requires providers of all search services likely to be accessed by children to accept and take appropriate action in response to complaints from users if their access to content is restricted because of an incorrect assessment of their age. We discuss this further below.

and interested persons from harm. A responsible person, team or function should be nominated to lead this triage process and ensure relevant complaints for services likely to be accessed by children reach the most relevant function or team.

b) relevant complaints for services likely to be accessed by children should be dealt with:

i) in a way that protects users and the provider's compliance with other applicable laws in question;

ii) within timeframes the provider has determined are appropriate; and

iii) in accordance with the other appropriate action recommendations set out above.

18.261 These recommendations largely codify the requirement in the Act. Where we make different recommendations for providers of services of different sizes or risk profiles, this is because we consider the minimum level of appropriate action is different depending on the type and level of risk of a service. This is because of the different volumes and types of complaints such services are likely to receive.

18.262 Other than a few minor changes to the wording we are suggesting above, these proposals mirror equivalent ones in the draft Illegal Content Codes, which recommend all providers of search services take appropriate action in response to other types of complaints, such as complaints about suspected illegal content. We have designed this proposal so that providers who should apply both measures may operate a single complaints process to provide for appropriate action in response to various different types of complaints, should they wish to do so.

Justification for the measures

18.263 For the most part, the analysis below does not differ substantially from the discussion in Section 16, Reporting and complaints, of our Illegal Harms Consultation. However, we have updated our language in places to clarify our recommendations and make our rationale easier to follow. In some places we have also changed our reasoning to make it pertinent to relevant complaints for services likely to be accessed by children, as set out in section 21(5) of the Act.

Measure UR5 (a): complaints about content considered harmful to children

18.264 Once a complaint has been received, it should enter the provider's search moderation function. This means providers will need to handle the complaint in accordance with Search Moderation Measures SM1 and SM2 in Section 17.

18.265 The Act sets out categories of content harmful to children, in relation to which providers must use proportionate systems and processes designed to minimise the risk of children encountering in search result PPC, PC and NDC.[571] Providers have a choice about how they assess content to meet these duties. If providers use different categories of content in their publicly available statements from those used in the Act, but they are nonetheless confident that their alternative categorisation secures the same protections for children as required by the Act, then they may assess complained about content using the categories in their publicly available statements. If the categories in their publicly available statements do not secure the protections for children required by the Act, then the provider will need to assess complained about content using the categories defined in the Act and explained in Volume 3, Section 8, Ofcom's Guidance on Content Harmful to Children.

[571] Section 29(3) of the Act.

18.266 As set out in Section 17, providers of large general search services and services that are multi-risk for content harmful to children (including multi-risk vertical search services) should prioritise the complaint in accordance with their prioritisation process and performance targets.

18.267 Providers of smaller services which are low-risk or single-risk for content harmful to children, which may include general search services and vertical search services, may not receive many, if any, complaints across diverse types of potentially harmful content, and may therefore not require prioritisation processes. We are therefore not recommending that these services must establish prioritisation processes and performance targets. We consider that if a provider of a smaller and low-risk or single-risk service has elected to establish a prioritisation process and performance targets for itself, it would be appropriate to abide by them. But a provider which has none would need to process all complaints received promptly.

18.268 If the content was determined by the provider to be either content harmful to children or content of a kind covered by the publicly available statement, the service provider would then need to take steps to provide for compliance with the safety duties protecting children.

18.269 As explained in Section 9 of Volume 3, we recognise that some downstream general search services may not be in control of the operations of the search engine. In such a case, we expect the upstream search service would be the provider of the search service and would need to secure compliance with the complaints handling duty. However, there may be circumstances in which the downstream entity does exercise control, and in those circumstances the downstream service would be the provider. We consider that this measure should apply to them similarly, since they can secure by contract that complaints are dealt with appropriately.

Measure UR5 (b): appeals

18.270 The Act requires providers to enable website owners to complain if their content is wrongfully downranked or no longer appears in search results (for example, as a result of filtering, deindexing or delisting) for being content harmful to children. We refer to these types of complaints as 'appeals.'

18.271 There are a number of different technical measures that might impact the visibility of search content in the manner contemplated by this duty. For example, content may "no longer [appear] in search results" following deindexing (which involves the removal of URLs or domains from a search index such that the URLs are prevented from appearing in search results), filtering (which involves ensuring that content is not returned in search results based on whether a condition is/is not met. For example, 'not displaying search results where condition "PPC" is true.') and possibly other actions. The primary technical means of which Ofcom is aware that might result in content being given a lower priority is "downranking" (which involves altering the ranking algorithm to ensure that a particular piece of content appears lower in the search results and is, therefore, less discoverable to users), which we use throughout this section for brevity. We note that our proposed Search Moderation Measures SM1A and SM1B recommend that providers of certain search services blur or consider whether it would be appropriate to blur (as relevant), image-based search content that is harmful to children. We do not consider that this action would fit within the scope of this duty, and therefore we would not expect service providers to handle complaints about blurring applied to search content.

18.272 Some providers may choose to run appeals through their main search moderation function. Others may establish a separate team. In either case, questions arise for providers about how quickly it is appropriate to review the decision, and what priority to give it as against other decisions.

18.273 For providers of large services (apart from vertical search services) and of services that are multi-risk for content harmful to children (including multi-risk vertical search services), we consider the volumes of appeals they are likely to need to consider, possibly across different types of potentially harmful content, are such that website owners may be harmed if they do not consider appropriate prioritisation in advance. We provisionally consider that providers of large services that are not vertical search services and services that are multi-risk should have regard to the matters set out under Measure UR5 (b) (i) in the 'Explanation of the measure' section above in determining what priority to give to review of the appeal.

18.274 For providers of services that are smaller and low-risk or single-risk and of vertical search services that are large and low-risk or single-risk, we provisionally think that there is no need to make detailed recommendations in Codes on prioritisation. We set out our reasons for this in Section 17, Search Moderation.

18.275 On the timeliness of considering appeals, for all the reasons set out in Section 17, we do not consider it appropriate for Ofcom to make specific recommendations. For providers of services which are smaller and low-risk or single-risk and large, low-risk or single-risk vertical search services, which we expect will not receive many complaints, let alone many appeals, we consider it will be sufficient to say that appeals should be determined promptly.

18.276 However, we consider that taking this approach for providers of large services (other than vertical search services) and services that are multi-risk for content harmful to children (including multi-risk vertical search services) could create perverse incentives and lead to harm, for example by incentivising providers to resolve complaints quickly rather than accurately. We therefore propose to recommend that such services should include in their content policies, targets as to speed and accuracy for the determination of appeals.[572] Our reasoning for this is the same as set out in relation to content moderation decisions in Section 17, Search Moderation. Similar recommendations in Section 17 as to monitoring and resourcing would apply in relation to these too, for the reasoning given there.

18.277 We consider that if, on review, a service reverses a decision that a URL or database contained content harmful to children, the service should:

a) reverse the action taken against the website owner or in relation to the content (or both) as a result of that decision (so far as appropriate for the purpose of restoring the position to what it would have been had the decision not been made);
b) where necessary to avoid similar errors in future, adjust the relevant content moderation policies; and
c) where necessary to avoid similar errors in future, take such steps as are within its power to secure that the use of automated moderation technology does not cause the same piece of content to be filtered or deprioritised again.

[572] There are a number of different ways providers could monitor performance against accuracy targets. For example, providers might want to select a sample of appeal decisions for a second review and track the number of decisions that were overturned by a second reviewer. We do not propose to be prescriptive about how providers should monitor performance against accuracy targets.

18.278 The policy intention behind point (a) here is that the provider should reverse the action they took against the website owner or the content, for example by removing any restriction placed on it if it had been downranked or reinstating it if it had been deindexed. We recognise that it may not be practical to restore the content to the exact position it would have been in had it not been incorrectly judged to be content harmful to children (e.g., the same position in search results) and this is not the intention of this recommendation.

Measure UR5 (c): complaints about incorrect assessments of age

18.279 We are not at this time proposing to recommend that providers of search services should operate age assurance to meet their safety duties protecting children (see Section 17 Search Moderation). However, providers may nonetheless choose to do so, and the Act requires providers of all search services likely to be accessed by children to accept, and take appropriate action in response to, complaints from users if their access to content is restricted because of an incorrect assessment of their age.

18.280 We are not aware of any search providers who currently operate age assurance and consider that search providers are unlikely to receive many complaints of this nature. Measure SD2 in Section 17 proposes that large search services apply safe search settings for users believed to be children and to ensure they cannot turn these settings off. It is possible that search providers may receive a higher number of complaints if adult users are incorrectly brought into scope of the safe search settings under Measure SD2.

18.281 As it is difficult to estimate the impact of Measure SD2 on complaints about incorrect assessment of age submitted to search providers, we provisionally are not making additional recommendations on prioritisation. Rather, we think it is sufficient to recommend that all such complaints should be handled promptly.

18.282 We consider that if, on review, a provider reverses a decision to restrict a user's access to content on the basis of an incorrect assessment of their age, the provider should restore the user's ability to access content on the service to an equivalent position to the one it would have been in, had the assessment of age been correct.

Measure UR5 (d): complaints about non-compliance with the safety duties protecting children

18.283 The Act also requires services to enable users and affected persons to complain if they consider the provider is not complying with a safety duty protecting children. We note that there is a significant risk of overlap between complaints about compliance with the safety duties protecting children and complaints about search content harmful to children, or about actions (such as filtering or downranking) being applied to content incorrectly. Where a complaint falls into one of those categories as well as this, we provisionally consider it appropriate for the provider to handle it in accordance with our proposed recommendations for those complaint types. However, we do not think we need to specify this in a specific measure, since this is already captured by the measures discussed above.

18.284 We provisionally think the appropriate action for providers in relation to complaints concerning compliance with safety duties protecting children would be to establish a triage process aimed at protecting users, affected persons and website owners from harm, including harm to their rights, such as to freedom of expression and privacy. A responsible person, team or function for such complaints should be nominated to lead this triage process and ensure complaints reach the most relevant function or team. They should be dealt with in a way that protects users and the provider's compliance with other applicable

laws in question, within timeframes the provider has determined are appropriate, and in accordance with our other proposed Code measures relating to complaints.

18.285 At this stage, we are not in a position to predict with sufficient certainty the many different types of complaint that may be submitted to providers relating to compliance with the safety duties protecting children, or to set out what action would be appropriate in response to them. Consequently, we are not currently proposing to make detailed recommendations in Codes about what final action may be appropriate for these complaints, although we will keep this position under review.

Rights assessment

Freedom of expression

18.286 We consider that the impacts of this measure would be very similar to those set out in Measure UR4 above. Given the clear parallels with that measure proposed for U2U service providers, please refer to our Rights Assessment for Measure UR4 to see our reasoning for our assessment of the impacts on rights in respect of measures applied to search services.

18.287 For the reasons set out in UR4, our provisional conclusion is that we do not think there would be a negative impact on service providers' rights to freedom of expression by proposing recommendations that set out what we think appropriate action will entail. We have designed this proposed measure with flexibility that allows service providers to decide how they implement it and what the outcomes of complaints would be, provided outcomes are in line with the Act's objectives to protect children from content harmful to them.[573] We think that there would be a positive impact on service providers' rights to freedom of expression by recommending they set out clear processes and providing them with the flexibility to determine their own boundaries, so long as they comply with the requirements of the children's safety duties set out in the Act.

Privacy

18.288 We do not consider that our proposed measure recommending that service providers take appropriate action in relation to complaints should have a negative impact on users' (including children and adults) rights to privacy in addition to those we have set out above in Measures UR1, UR2 and UR4.

18.289 We have set out our rationale for this in Measure UR4 above, which we think will also apply to this proposed measure for search services. Given the clear parallels with that measure proposed for U2U service providers, please refer to our Rights Assessment for Measure UR4 to see our reasoning for our assessment of the impacts on rights in respect of measures applied to search services.

18.290 For the reasons set out in Measure UR4 above, we therefore consider that the impact of the proposed measure on individuals' (including adults' and children's) rights to privacy to be relatively limited, and potentially overall positive. It is likely to constitute the minimum degree of interference required to secure that service providers fulfil their children's safety duties under the Act. Taking this, and the benefits to children into consideration, we consider that it is therefore proportionate.

[573] Section 29 of the Act.

Impacts on services

18.291 The costs of taking appropriate action for complaints will vary across different types and sizes of services, and for services with different levels of risk. While we expect the costs could be very significant for some service providers, we believe they derive, in large part, from duties in the Act.

18.292 We are proposing to mitigate the risk of imposing unnecessary costs of our recommendations by allowing service providers flexibility to set their own timescales for resolving complaints. This will help to ensure that the costs incurred are proportionate to the nature and risk profile of the service.

18.293 It should also be noted that if complaints about content harmful to children are routed through a service provider's content moderation function, the costs of taking appropriate action in the case could be regarded as part of search moderation.

18.294 Furthermore, while we recognise that it depends in part on the nature of the service, we would generally expect the volume of complaints a service provider receives to increase with the size of the service, the risks on the service, the volume of search queries users run, and the number of search moderation decisions being taken by the provider. This means that the highest costs will be incurred by the providers of the largest services who are most likely to be able to absorb them, and who we expect would see the greatest benefits from implementing these recommendations.

18.295 The costs of establishing and running a triage process to ensure that relevant complaints for services likely to be accessed by children reach the most relevant function will also depend on the nature of the service, but we expect these costs to scale with the size and complexity of a service, and the volume of complaints.

18.296 Services whose current complaints policies do not meet the measure will also incur costs when adapting their policies regarding what action they will take in response to complaints, and ensuring they are compliant. Services with more complex governance processes are likely to incur greater costs when agreeing these policies.

18.297 To implement Measure UR5 (b), a service would need to develop a prioritisation framework and set and monitor performance targets for appeals. We believe that this would involve similar activities and therefore costs as those described in the context of Search Moderation Measures SM3 and SM4. We consider that there are likely to be some overlaps in the processes required for SM3/SM4 and the measure UR5 (b) proposed here, which may imply some cost savings for services.

18.298 We are not at this time proposing to recommend that providers of search services should operate age assurance to meet their safety duties protecting children. However, we acknowledge that large search services may get some additional complaints related to incorrect assessments of age as a result of our proposed measure SD2. More generally, if a service provider chooses to implement age assurance, they will potentially have to handle a large volume of complaints, which could lead to significant costs. The volume of complaints about incorrect assessment of age a service receives will tend to vary with the volume of decisions it makes. This means that, if service providers choose to implement age assurance, the services with the highest costs will tend to be large services, with the greatest ability to bear those costs.

18.299 We recognise that all service providers who should apply measure UR5 should also apply the related proposed measure in the Illegal Harms consultation.[574] We consider that there will be scope for significant cost savings where service providers use the same systems and processes to provide for appropriate action in relation to the types of complaints covered by the draft Illegal Content Codes and the draft Children's Safety Codes.

18.300 Additionally, we have considered the potential added complexity for all kinds of services in making judgements about content harmful to children. However, as set out above, the Act does not necessarily require services to make judgements about content harmful to children if they are satisfied that their publicly available statements already include provisions that ensure that content that would be considered content harmful to children under the Act is appropriately dealt with in line with their duties. To the extent that new judgements about content harmful to children are required, this is down to the requirements of the Act.

Which providers we propose should implement this measure

18.301 Due to the fact that the reporting and complaints duties apply to all providers of in-scope services likely to be accessed by children, we have proposed setting out broad features (as opposed to specific ones) that we recommend providers consider when designing their reporting and complaints processes. We believe we have approached this in a way that seeks as far as possible to elucidate the basic legal requirements set out in the Act.

18.302 On this basis, we believe our proposals regarding complaints about content harmful to children (Measure UR5 (a)), complaints about incorrect assessment of a user's age (Measure UR5 (c)) and complaints about non-compliance with the safety duties protecting children (Measure UR5 (d)), are proportionate and suitable for providers of all search services likely to be accessed by children.

18.303 For appeals Measure UR5 (b), while we propose that this would apply to all search services, we propose to make different recommendations for providers of large services (other than vertical search services) and services that are multi-risk for content harmful to children (including multi-risk vertical search services) compared to providers of other services. This is because these types of services are likely to receive a high volume of these types of complaints across a range of different types of content that may be harmful to children. We consider that unless providers of these services consider prioritisation and performance targets in advance, there is a risk that they will be unable to take appropriate action in response to large volumes of complaints.

18.304 We consider that the benefits of adopting a prioritisation framework and setting and monitoring performance targets for providers of these types of large services (other than vertical search services) or risky services are sufficiently important for them to incur the costs of doing so, in order to be able to take appropriate action in response to complaints, given the larger volume of complaints likely to be present on such services.

18.305 Providers of services that are neither large nor multi-risk for content harmful to children are less likely to receive large volumes of complaints across a diverse set of content types that may be harmful to children. As such we do not think it is necessary to recommend that they should establish prioritisation processes or set and monitor performance targets. Given there are likely to be costs involved in implementing these, we do not believe that the potential benefits are large enough to justify these costs to such services. Rather, we think it

[574] Our Illegal Harms Consultation Volume 4, Section 16, Measure 4.

is sufficient and proportionate to recommend that providers of these services should handle these complaints promptly. We consider this is the minimum necessary for providers of these services to comply with their duty under the Act.

18.306 See Section 17, Search Moderation, for further discussion of this approach.

Provisional conclusion

18.307 Given the harms this measure seeks to mitigate in respect of content harmful to children, as well as service providers' duties to operate processes that provide for appropriate action to be taken in response to complaints about content harmful to children and other types of complaints, we consider this measure appropriate and proportionate to recommend for inclusion in the draft Children's Safety Codes. For the draft legal text for this measure, please see PCS C5-C8, C10 and C11 in Annex A8.

Further measures considered

18.308 In response to our 2023 CFE, stakeholders suggested we recommend a number of other measures. We have discussed some of these earlier in this section. Here we set out our thinking in relation to some further suggestions not covered in the discussion of our proposals above.

18.309 Several stakeholders called for services to provide additional support to children during or after reporting harmful content.[575] For instance, one stakeholder called for services to operate a dedicated helpline to support people reporting.[576] Another called for services to provide a specialist mental health support team.[577] As discussed above, we are proposing to recommend two measures that would make it easier for children to report harmful content and support children following exposure to certain types of content harmful to children. Measure UR2 sets out how services can ensure their complaints processes are easy for children to find, use and access. User Support Measure US5 in Section 21 recommends that service providers should signpost children who report certain types of content to appropriate support resources, including resources about mental health. We consider that these measures are more proportionate and less prescriptive ways to ensure children are supported to report potentially harmful content.

18.310 Another stakeholder, Papyrus, recommended that services should notify an adult when children report harmful content.[578] We do not currently have evidence for the effectiveness of this suggestion as a measure to protect children from harmful content. We also have concerns that it could have a negative impact on children's right to privacy, and potentially discourage children from reporting, since we know children sometimes worry about getting in trouble for viewing harmful content. We therefore do not propose to recommend this measure at this stage. However, we may reconsider this in future should new evidence come to light.

18.311 We have also considered other potential measures relating to complaints processes. However, we are not currently in a position to propose they are included in the draft

[575] Samaritans response to 2023 Protection of Children Call for Evidence.
[576] Girlguiding response to 2023 Protection of Children Call for Evidence.
[577] Mental Health Foundation response to 2023 Protection of Children Call for Evidence.
[578] Papyrus response to 2023 Protection of Children Call for Evidence.

Children's Safety Codes because of lack of evidence for their effectiveness and costs at this early stage of the regime. We plan to continue gathering evidence to inform our future work, for example through research and using our formal information gathering powers where appropriate. If we can address the current evidence gaps, we may reconsider whether to propose these potential measures for future iterations of the Children's Safety Codes:

a) **Trusted flaggers programmes:** We considered recommending that providers establish trusted flaggers programmes to enable third-party organisations to make priority reports of harmful content. Third-party experts may be well placed to help identify emerging trends in content harmful to children. Evidence suggests that trusted flaggers tend to provide more accurate reports than general user reports.[579] Using trusted flaggers would also take the onus off children to report harmful content themselves. While some service providers currently use trusted flaggers for some illegal content, and we proposed a dedicated reporting channel for trusted flaggers of fraudulent content in our draft Illegal Content Codes, we do not have sufficient evidence on the effectiveness or cost of these programmes for content harmful to children.
b) **Communicating the outcome of complaints to users:** As we discuss further above, we considered recommending that providers should communicate the outcome of complaints to complainants, particularly children. However, before proposing such a measure we would require further evidence of the practicalities and costs of implementing such a requirement at scale. See Measure UR3 above for a more detailed discussion of our considerations.
c) **Dedicated reporting channels:** We considered recommending service providers create dedicated reporting channels for children or types of content harmful to children as defined in the Act. While several service providers currently operate such channels for certain types of content, and we proposed a dedicated reporting channel for fraud (to be used by trusted flaggers) in our draft Illegal Content Codes, we do not currently have evidence for the effectiveness of reporting channels dedicated to reports made by children or about content harmful to children as defined in the Act. We provisionally consider that the content moderation proposals in Section 16, Content moderation for U2U services, about how providers of large services should prioritise content for review, will help ensure reports submitted by children and about content harmful to children as defined in the Act are prioritised appropriately. However, we may consider whether to recommend dedicated reporting channels for specific types of content as part of our future work.

18.312 We would welcome any evidence stakeholders can provide for the effectiveness, costs, and risks of these potential future measures.

[579] EU Directorate-General for Justice and Consumers, 2021. Countering illegal hate speech online: 6th evaluation of the code of conduct [accessed December 2023]; Clare Lilley, EMEA lead on Child Safety, Google. Oral evidence to the Science and Technology Committee, Tuesday 16 October 2018, Q484 [accessed December 2023].

19. Terms of service and publicly available statements

Terms of service ('terms') and publicly available statements ('statements') typically lay out the rights and responsibilities that a service provider and the users of their service have towards one another.

The Act places duties on all service providers regarding the substance and presentation of terms and statements. In our Illegal Harms Consultation, we proposed two measures to help providers meet terms and statements duties regarding illegal content on all services (see Chapter 17, Terms of service and publicly available statements). Following the publication of our categorisation advice to the Secretary of State in March 2024, we are proposing one additional measure for inclusion in our draft Illegal Content Codes, recommending that providers of Category 1 and 2A services meet their additional terms and statements duties regarding illegal content (Measure 6AA). We consult on this measure below.

In this consultation, we are proposing three measures to help providers of services likely to be accessed by children to meet their duties relating to terms and statements that will build on the measures proposed for inclusion in our draft Illegal Content Codes.

Children and the adults who care for them must refer to terms or statements if they want to understand how service providers keep children safe while using their service. If this information is not provided, or if it is presented in a confusing or inaccessible way, children, parents and other carers may not be able to make an informed choice about whether to use a service. Moreover, when using the service, they may not understand their rights and responsibilities as users, nor recognise content that is harmful to children and take action in response to it. This could contribute to the prolonged presence of content harmful to children on a service.

To address these risks to children, and in accordance with the children's safety duties in the Act, we are proposing measures targeting the substance and presentation of terms and statements relating to the protection of children on a service. We have assessed the potential impacts of our proposals, including costs and rights impacts, and deem them proportionate measures for all U2U and search services in scope of the children's safety duties.

Our proposals

#	Proposed measure	Who should implement this [580]
TS1	**Terms and Statements** regarding the protection of children **should contain all information mandated by the Act**	All Search and U2U services
TS2	**Terms and Statements** regarding the protection of children **should be clear and accessible**	All Search and U2U services
TS3	Services should **summarise the findings of their most recent children's risk assessment in their Terms or Statement**	All Category 1 and 2A services

[580] Proposed measures TS1, TS2 and TS3 relate to providers of services likely to be accessed by children.

Consultation questions

46. Do you agree with the proposed Terms of Service / Publicly Available Statements measures to be included in the Children's Safety Codes? Please confirm which proposed measures your views relate to and provide any arguments and supporting evidence. *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.*
47. Can you identify any further characteristics that may improve the clarity and accessibility of terms and statements for children?
48. Do you agree with the proposed addition of Measure 6AA to the Illegal Content Codes? Please provide any arguments and supporting evidence.

Why are terms and statements important for protecting children?

Definition Box 1: Defining terms and statements

The Act connects terms to U2U and combined services, and statements to search services.

- It defines terms of service, in relation to U2U services, as "all documents (whatever they are called) comprising the contract for use of the service (or of part of it) by United Kingdom users".[581]
- It requires search services to produce a statement setting out certain information about how they operate. This statement must be made "available to members of the public in the United Kingdom".[582]
- It permits combined services, which have both U2U and search functionalities, to set out what would be required in a publicly available statement in terms of service instead.[583]

19.1 Terms and statements contain information about how a service functions, including who is allowed to use the service, rules for use of the service and how users will be protected from harm on the service.

19.2 Where terms and statements lack information regarding the protection of children, or where this information is presented in a way that is confusing or inaccessible to children, this can present risks to children using the service.[584]

19.3 It is therefore important that services likely to be accessed by children have terms and statements that are clear and accessible to children. This will help children, independently or in consultation with adults who care for them, to:

a) make an informed choice about whether to use a service;
b) understand the measures a service uses to keep them safe from harmful content, including whether, when and how children can control their online experience;
c) understand how a service handles complaints procedures if something goes wrong.

19.4 As a result, children should have knowledge of, and confidence in, the services that they use. Understanding how a service intends to keep them safe from content harmful to children, including any means the service provides for them to control their own user experience, should help children to recognise and take action if they are exposed to harmful content online. This should contribute to a safer online environment for children.

[581] Section 236 of the Online Safety Act 2023.
[582] Definition of 'publicly available' taken from section 236 of the Act.
[583] Section 25(2)(a) of the Act.
[584] For evidence on why accessible terms of service are important for children, see 'Governance, Systems and Processes' in the Ofcom Children's Register of Risks at Section 7.11 in Volume 3.

What are regulated services' obligations regarding terms and statements?

19.5 For both illegal content and content that is harmful to children, the Act's duties relating to provisions in terms and statements may be grouped under three core areas:

- Substance;[585]
- Consistency;[586] and
- Clarity and accessibility.[587]

19.6 Our proposals for the protection of children deal with all three areas. Measures TS1 and TS3 address substance, while Measure TS2 addresses clarity and accessibility. Measure TS1 also addresses consistency, but only regarding the duty for U2U service providers to consistently apply any provisions in their terms of service detailing any measures they use to prevent access to their service by children under a certain age.[588] This consistency duty is an explicit requirement for our code of practice for the U2U children's safety duties.[589] There are also duties for U2U and search services to consistently apply provisions in their terms or statement explaining how children are to be prevented or protected from encountering harmful content.[590] The Act lays out equivalent duties regarding illegal content,[591] so here we adopt the same approach as we did in our Illegal Harms Consultation. In our view, providers who properly implement our recommendations to prevent or protect children from encountering harmful content will necessarily do so in a way that ensures terms or statements are applied consistently (by virtue of how those recommendations have been designed).

19.7 Outside of duties regarding illegal content and content that is harmful to children, there are other duties in the Act relevant to terms and statements, including the "additional terms of service duties".[592] We will consult on any proposals relating to these duties in the categorised services consultation, due to be published in early 2025.

19.8 The measures described in this section are compatible with the pursuit of the online safety objectives laid out in the Act, in particular that "United Kingdom users (including children) are made aware of, and can understand, [terms and statements]".[593]

19.9 However, we recognise that no matter how clear and accessible they are, some children might not be able to fully understand information in written terms and statements.[594] While steps can be taken to make these documents clearer and more accessible, they are contractual in nature and do not easily lend themselves to being child friendly.

---

[585] Regarding the protection of children, see sections 12(9), 12(11)(a), 12(12), 12(14), 21(3), 29(5), 29(7), 29(9) and 32(3) of the Act.
[586] Regarding the protection of children, see sections 12(10), 12(11)(b) and 29(6) of the Act.
[587] Regarding the protection of children, see sections 12(13), 29(8), 21(3), 32(3) and Schedule 4, paragraph 4 (a)(iii) and paragraph 5 (a)(iii) of the Act.
[588] This consistency duty is set out in section 12(11)(b) of the Act.
[589] Schedule 4, paragraph 6(a) of the Act.
[590] Sections 12(10) and 29(6) of the Act.
[591] Sections 10(6) and 27(6) of the Act.
[592] Sections 71 and 72 of the Act.
[593] Schedule 4, paragraph 4 (a)(iii) and paragraph 5 (a)(iii) of the Act.
[594] This might particularly be true of young children and those with poor reading skills or learning difficulties.

19.10 We are therefore proposing to make an additional recommendation – Measure US6 in Section 21, User Support – on the need for age-appropriate user support materials. This proposed measure recommends that in-scope service providers create visual, audio-visual, or interactive materials for children, and guidance for the adults who care for them, to explain the user tools and reporting and complaints functions available to help children control their experience on a service. These materials should be specifically designed to help children understand the proactive steps they can take to feel safer online. This should ensure that important information about staying safe from harmful content is widely accessible to children in formats that are easy for them to comprehend, even if they cannot fully understand this information within terms and statements.

Interaction with Illegal Harms

19.11 In our Illegal Harms Consultation we proposed the following measures regarding terms and statements to be included in our draft Illegal Content Codes:

a) **Measure 6A**: All U2U and search service providers should include provisions in their terms or statements regarding the protection of individuals from illegal content, any proactive technology used, and information on how complaints are handled and resolved.
b) **Measure 6B**: All U2U and search service providers should ensure that relevant provisions included in terms or statements regarding the protection of individuals from illegal content are clear and accessible.

19.12 Refer to Section 17 of our Illegal Harms Consultation, Terms of service and publicly available statements, for a detailed discussion of the evidence, costs and impacts of these measures.[595]

19.13 We understand that many service providers produce just one version of their terms or statement. This one version may explain their approach to keeping all users safe from illegal content, as well as their approach to keeping children safe from content that is harmful to them. We have taken this into account when assessing the impact of recommending measures for inclusion in our draft Children's Safety Codes.

19.14 We provisionally consider that proposed Measures 6A and 6B in the draft Illegal Content Codes are also proportionate for providers of services likely to be accessed by children in relation to their additional terms and statements duties for the protection of children. We set out below our detailed assessment of the evidence and impacts of these measures as they relate to these duties.

19.15 Proposed Measure TS1 for the Children's Safety Codes is slightly amended from the equivalent provisional recommendation in the draft Illegal Content Codes (Measure 6A). This reflects the differing provisions that must be included in terms and statements in relation to the protection of children, particularly the duty on U2U services to consistently apply any measures they take to prevent access to their service by children under a certain age.

19.16 Proposed Measure TS2 is in substance unchanged from our equivalent provisional recommendation in the draft Illegal Content Codes (Measure 6B), although we have, where relevant, relied upon updated evidence and language to consider the specific rationale as to

[595] Ofcom, 2023. Protecting people from illegal harms online. Volume 4: How to mitigate the risk of illegal harms – the illegal content Codes of Practice.

why we propose this measure to be relevant for services likely to be accessed by children and their duties under the Act.

19.17 We are proposing to recommend that service providers write their terms or statement to a reading age comprehensible for the youngest person permitted to use the service without consent from a parent or guardian. This language is different from measure 6B in our draft Illegal Content Codes, where we recommended that service providers write their terms or statement to a reading age comprehensible for the youngest person permitted to agree to them.

19.18 Due to the close similarities of our proposed Measure TS2 and measure 6B, we are consulting on updating this language for both the draft Children's Safety Codes and the draft Illegal Content Codes. Given responses to our Illegal Harms Consultation, we also consider that this updated language will better clarify our recommendation around the reading age that terms or statements should be written to.

19.19 We believe the updated language better reflects the way many search and U2U services require parental or guardian consent for children under a certain age to agree to terms or statements and the fact that, despite best efforts to draft terms or statements simply and clearly, children will often need support to understand the service's public facing information. This updated language, however, does not change or shift the onus that is on service providers to make available clear and accessible terms or statements that will empower children to independently, and/or with the adults who care for them, have safer experiences online.

19.20 Following the publication of our categorisation advice to the Secretary of State in March 2024,[596] we are proposing the inclusion of Measure TS3 in our draft Children's Safety Codes and an equivalent measure for inclusion in our Illegal Content Codes (measure 6AA). In both cases, the proposed measure codifies the additional duty under the Act on providers of Category 1 and 2A services to summarise the findings of their most recent illegal content risk assessment or children's risk assessment in their terms or statement.[597]

Our proposals to protect children

19.21 The Act requires all providers of U2U and search services likely to be accessed by children to explain in their terms or statement the details of certain provisions taken to keep children safe on their service.[598] This information must be clear and/or accessible to users, including children.[599]

19.22 We are proposing two measures to help all in-scope service providers meet their duties in this area, ensuring that users, including children, can better access and understand reliable information about safety practices on in-scope services:

a) **Measure TS1:** Services should ensure that provisions included in terms or statements regarding the protection of children contain all the information mandated by the Act. U2U services must additionally ensure that they consistently apply provisions in their

[596] Ofcom, 2024. Categorisation: Advice Submitted to the Secretary of State. Subsequent references are to this document throughout.
[597] See sections 12(14) and 29(9), 10(9) and 27(9) of the Act for full details of these duties.
[598] Sections 12(9), 12(11)(a), 12(12), 21(3), 29(5), 29(7) and 32(3) of the Act.
[599] Sections 12(13), 29(8), 21(3), 32(3) and schedule 4, paragraph 4 (a)(iii) and paragraph 5(a)(iii) of the Act.

terms detailing any measures they use to prevent access to their service by children under a certain age.

b) **Measure TS2:** Services should ensure that relevant provisions included in terms or statements regarding the protection of children are clear and accessible.

19.23 We are further proposing an additional measure to help Category 1 and 2A service providers meet a duty applying only to them:

a) **Measure TS3 (Children's Safety Code):** Category 1 and 2A services that are likely to be accessed by children should summarise the findings of their most recent children's risk assessment in their terms or statement.[600]

b) **New Measure 6AA (Illegal Content Code):** Category 1 and 2A services should summarise the findings of their most recent illegal content risk assessment in their terms or statement.[601]

19.24 In the rest of this section, we set out our rationale for these proposals. As discussed below, Measure TS1 and Measure TS3 codify specific requirements within the Act. We set out our rationale for Measure TS2 in more detail, including which services we propose this measure applies to.

Measure TS1: Terms and statements regarding the protection of children contain all information mandated by the Act

Explanation of the measure

19.25 In delivering this measure, we would expect to see providers of U2U and search services likely to be accessed by children develop or revise their terms or statement, ensuring they include provisions mandated by the Act that relate to the protection of children. The measure further states that providers of U2U services likely to be accessed by children should consistently apply any provisions in their terms detailing any measures they take to prevent access to the service by children under a certain age.

19.26 This measure mirrors an equivalent measure in the draft Illegal Content Codes (Measure 6A), which recommends all U2U and search service providers should include provisions in their terms or statements regarding the protection of individuals from illegal content, any proactive technology used, and information on how complaints are handled and resolved. Providers in scope of both measures may operate a single version of their terms or statement to cover both measures if they wish.

19.27 Table 19.1 below details the Act-mandated provisions that U2U and search services must include in their terms or statement with regard to the protection of children.

[600] Section 12(14) and 29(9) of the Act.
[601] Section 10(9) and 27(9) of the Act.

Table 19.1: Duties on providers of U2U and search services likely to be accessed by children to include mandated provisions relating to the protection of children in their terms/statement

	U2U service providers must include provisions in their terms of service specifying…	Search service providers must include provisions in their publicly available statement specifying…
1	Information about any **proactive technology** the service uses to safeguard children in line with their children's safety duties, including the kind of technology, when it is used, and how it works.[602]	
2	The policies and processes that govern the **handling and resolution of complaints** "of a relevant kind".[603]	
3	• How children of any age are to be prevented from encountering each kind of **primary priority content (PPC)** that is harmful to children. • How children in age groups judged to be at risk of harm from **priority content (PC)** that is harmful to children are to be protected (where they are not prevented) from encountering each kind of PC. • How children in age groups judged to be at risk of harm **from non-designated content (NDC)** that is harmful to children are to be protected (where they are not prevented) from encountering each kind of NDC.[604]	How children are to be protected from: • each kind of **primary priority search content (PPC)** that is harmful to children; • each kind of **priority search content (PC)** that is harmful to children; and • non-designated search content (NDC) that is harmful to children.[605]
4	Details about the operation of any measure taken or used by the service that is designed to **prevent access** to the whole, or part of, the service **by children under a certain age**.[606]	

[602] See sections 12(12) and 29(7) of the Act for full details of these duties. Refer to Annex 15 for a definition of proactive technology.
[603] See sections 21(3) and 32(3) of the Act for full details of these duties. The list of complaints of a relevant kind for services likely to be accessed by children are set out in sections 21(5) and 32(5) of the Act. Refer to Section 18 of this volume, User reporting and complaints, for our recommendations relating to the handling and resolution of complaints.
[604] See section 12(9) of the Act for full details of these duties. PPC, PC and NDC are respectively defined in sections 60, 61 and 62 of the Act.
[605] See section 29(5) of the Act for full details of these duties.
[606] See section 12(11)(a) of the Act for full details of this duty.

19.28 In-scope service providers are likely to take different approaches to protecting children on their service. These approaches will determine the content they include in their terms or statement with respect to the duties in Table 19.1, above.

19.29 The duties in row 3 require in-scope service providers to explain in terms and statements how they will prevent or protect children from encountering content that is harmful to them. When presenting this information, service providers must be sure that users, including children, can understand how their service will protect children from each individual kind of PPC, PC and NDC. Where service providers use the same measure to protect children from multiple kinds of content, they need not repeat their explanation of that measure in their terms or statement, so long as it remains clear for each kind of PPC, PC and NDC which measures are being used, and how, to protect or prevent children from encountering it.

19.30 We expect that where in-scope service providers apply measures to meet the children's safety duties under the Act, they will in effect consistently apply to all users the provisions in their terms or statements specifying how children will be protected or prevented from encountering content that is harmful to them.[607] We are not proposing to make recommendations elsewhere in the Code as to how service providers might implement measures designed to prevent access by children under a certain age, for reasons discussed in Section 14 (Age Assurance).[608] Consequently, we are recommending that where services detail such measures in their terms or statement, they meet both the substance and consistency requirements under the Act.[609] This means that to comply with the Codes, in-scope service providers must include provisions in their terms or statements detailing the operation of any measure designed to prevent access to the service by children under a certain age *and* apply those provisions consistently for all users.

19.31 We recognise that this measure may overlap with in-scope service providers' existing approach to protecting children (where such approaches already exist). Where this is not the case, Measure TS1 will ensure a higher level of protection for children.

19.32 We also recognise that in-scope service providers may wish to manage the level of detail in their terms or statement, both to mitigate the risk that bad actors are able to use the terms or statement to circumvent safety measures, and to ensure that their terms or statement are clear and accessible, including to children (as required by the Act).[610] We understand that excessive levels of detail in terms or statements recording how systems and processes are being used to protect children may require a disproportionate use of resources to keep up-to-date.

19.33 Some service providers may choose to provide their terms or statement via multiple documents.[611] In-scope service providers must ensure that all mandated provisions regarding the protection of children meet the clarity and accessibility standard required by

[607] The Act requires that service providers apply the duties in row 3 of Table 1 consistently for all users (see sections 12(10) and 29(6) of the Act.
[608] Our draft Children's Safety Code for U2U services recommends measures to prevent access to certain services or parts of a service by all children; see Measures AA1 and AA2 in Age Assurance, Section 15.
[609] Sections 12(11)(a) and 12(11)(b) of the Act.
[610] Sections 12(13), 29(8), 21(3), 32(3) and Schedule 4, paragraph 4 (a)(iii) and paragraph 5 (a)(iii) of the Act.
[611] For full definitions of terms and statements as dictated by the Act, see Definition Box 1: Defining terms and statements.

the Act,[612] regardless of the number of documents that constitute the provider's terms or statement, or where this information is located within their terms or statement.

Rights assessment

19.34 This measure recommends that services likely to be accessed by children should ensure that provisions included in terms or statements regarding the protection of children contain all the information mandated by the Act. U2U services must additionally ensure that they consistently apply provisions in their terms detailing any measures they use to prevent access to their service by children under a certain age.

19.35 We have carefully considered whether this proposed measure would constitute an interference with users' (both children and adults) or services' freedom of expression or association rights, or user's privacy rights. Our provisional conclusion is that it would not. This proposed measure is intended to capture the specific requirements for services in relation to terms of service and publicly available statements under the Act. Whilst this includes provisions which set out, for example, how a service will prevent or protect children from encountering content that is harmful to them, it does not require a service to take specific action in relation to content or personal data. We additionally consider that the provision of the specific types of information mandated by the Act and set out above, would be beneficial to users in that they would be consistently provided with information about how the service operates across a number of key areas relating to children's online safety, the use of proactive technology, user access and complaints. This may have positive impacts on users' - particularly children's - rights to freedom of expression and association, and also their rights to privacy in that it should also help them understand how a service operates to protect them from encountering content that might be harmful to them, and protect their personal data as they use and gain access to the service.

Impacts on services

19.36 Providers of services likely to be accessed by children who do not currently include provisions in their service's terms or statement that meet the relevant duties outlined above will need to add these provisions and incur the relevant costs. Since this measure reflects a direct requirement of the Act, any costs or impacts to services associated with this measure result directly from the duty in the Act. We have therefore not considered any costs or impacts to services associated with this measure as part of assessing the implications of this measure for services.[613]

19.37 We consider the requirements set out in the duties above are sufficiently clear for services to implement without further elaboration by Ofcom. We recognise that all service providers in scope of this measure would also be in scope of the equivalent measure proposed in our Illegal Harms consultation. While we expect services to incur incremental costs to meet the requirements of the current measure over and above the corresponding Illegal Harms measure, we have not considered these costs since the measure states a direct requirement of the Act.

[612] Sections 12(13), 29(8), 21(3), 32(3) and Schedule 4, paragraph 4 (a)(iii) and paragraph 5 (a)(iii) of the Act.
[613] Sections 12(9), 12(11)(a), 12(11)(b), 12(12), 21(3), 29(5), 29(7) and 32(3) of the Act.

Which providers we propose should implement this measure

19.38 This measure will apply to providers of all U2U and search services likely to be accessed by children, as the Act requires these services to include in their terms or statements the relevant provisions mentioned above.

Provisional conclusion

19.39 This measure seeks to mitigate the risk of children and the adults who care for them not understanding children's rights and responsibilities as users of a service. We consider this measure appropriate and proportionate to recommend for inclusion in the draft Children's Safety Codes. For the draft legal text for this measure, please see PCU D1 in Annex A7 and PCS D1 in Annex A8.

Measure TS2: Terms and statements regarding the protection of children are clear and accessible

Explanation of the measure

19.40 In delivering this measure, we would expect providers of all U2U and search services likely to be accessed by children to compile and present certain provisions within their terms or statement in a clear and accessible way. In particular, service providers should focus on making these provisions clear and accessible for children. To achieve this, service providers should have regard to the findability and usability of these provisions, as well as how they are laid out and formatted, and the language used to describe them.

19.41 Based on available evidence, our measure recommends that service providers should ensure that relevant provisions are:

- easy to find for both users and non-users of the service;
- laid out and formatted in a way that helps users, including children, to understand them;
- written to a reading age comprehensible for the youngest person permitted to use the service without consent from a parent or guardian;[614]
- compatible with assistive technologies.

19.42 This measure relates to duties in the Act concerning the clarity and/or accessibility of certain provisions set out in Measure TS1.[615] It is also intended to support the online safety

[614] Please refer to paragraphs 1.15-1.18 above for a discussion of our position on reading age.
[615] Section 12(13) requires U2U services to ensure that certain provisions in their terms of service (laid out in sections 12(9), 12(11) and 12(12) of the Act) are clear and accessible. Section 29(8) requires search services to ensure that certain provisions in their publicly available statement (laid out in sections 29(5) and 29(7) of the Act) are clear and accessible. Section 21(3) requires that U2U services ensure provisions in their terms of service specifying the policies and processes that govern the handling and resolution of relevant complaints are easily accessible, including to children. Section 32(3) requires that search services ensure the policies and processes that govern the handling and resolution of relevant complaints are publicly available and easily accessible, including to children.

objective that "United Kingdom users (including children) are made aware of, and can understand, terms and statements".[616]

19.43 This measure mirrors an equivalent measure in the draft Illegal Content Codes, which recommends all U2U and search service providers should ensure that relevant provisions included in terms or statements regarding the protection of individuals from illegal content are clear and accessible. Providers in scope of both measures may operate a single version of their terms or statement to cover both measures if they wish.

Effectiveness at addressing risks to children

19.44 If information about the protection of children is not provided in terms and statements, or if it is presented in a confusing or inaccessible way, users, including children, may not be able to make an informed choice about whether to use a service. Moreover, when using the service they may not understand their rights and responsibilities as users, making it hard for them to recognise content that is harmful to them and to take action in response to it. This could contribute to the prolonged presence of content harmful to children on a service.[617]

19.45 Terms and statements are often long, confusing and require advanced reading skills to understand,[618] meaning they are unsuitable for many users, especially children.[619] For example, an assessment of existing privacy policies (which serve a similar purpose to terms and statements) suggests these are rarely targeted at children.[620]

19.46 Ofcom research found that UK internet users (including 16- and 17-year-olds) rarely access terms and statements when visiting websites or apps.[621] There is evidence that adults who

---

[616] Schedule 4, paragraph 4 (a)(iii) and paragraph 5 (a)(iii) of the Act.

[617] For evidence on why accessible terms of service are important for children see 'Governance, Systems and Processes' in the Children's Register of Risk at Section 7.11 in Volume 3.

[618] Ofcom, 2023. Regulating Video-Sharing Platforms (VSPs). Our first 2023 report: What we've learnt about VSPs' user policies. Subsequent references are to this document throughout; Ibdah, D., Lachtar, N., Meenakshi Raparthi, S. & Bacha, A., 2021. "Why should I read the privacy policy, I just need the service": A study on attitudes and perceptions toward privacy policies, *IEEE Access*, 9. [accessed 16 April 2024]. 55% of surveyed users did not correctly understand what a privacy policy told them; Taloustutkimus Oy (Turja, T. & Sandqvist, S.), 2021. The use of digital services 2021: Summary report. [accessed 16 April 2024]. Only 44% of survey respondents felt they understood well the terms and conditions of different applications and services.

[619] See Schneble, C.O., Favaretto, M., Elger, B.S. & Shaw, D.M., 2021. Social media terms and conditions and informed consent from children: Ethical analysis, *JMR Pediatrics and Parenting*, 4 (2). [accessed 16 April 2024]. Subsequent references are to this research throughout. A thematic analysis of terms and conditions on 20 social media platforms and two mobile phone operating systems, which concluded 'terms and conditions are often too long and difficult to understand, especially for younger users.' See also Milkaite, I. & Lievens, E., 2020. Child-friendly transparency of data processing in the EU: from legal requirements to platform policies. *Journal of Children and Media*, 14 (1). [accessed 16 April 2024]; The Children's Commissioner for England, January 2017. Growing up digital: A report of the Growing Up Digital Taskforce. [accessed 16 April 2024]. Subsequent references are to this document throughout; Ofcom video-sharing platform guidance, 2021; Mental Health Foundation response to 2023 Protection of Children Call for Evidence; Anti-Bullying Alliance response to 2023 Protection of Children Call for Evidence.

[620] 5Rights Tick to agree, 2021. Looked at 123 privacy policies for websites likely to be accessed by children, only 9 of which (7%) had a specific policy targeted at children.

[621] Ofcom, 2023. Platform Terms and Accessibility. Question 1: Have you ever needed to access terms of service/ guidelines on social media? Note: Only 33% of 16–24-year-olds reported ever needing to access social media terms and conditions, decreasing to 19% for all respondents.

do read them spend limited time doing so.[622] Most people choose to accept terms and conditions without reading them[623] and many say they do not understand them.[624]

19.47 However, Ofcom research found that 29% of 16–24-year-olds would check a platform's community guidelines and 7% would check the terms and conditions if they were unsure about posting something on the platform.[625] It is therefore important that these documents contain clear and accessible information for users, including children, when they need it.

19.48 There is evidence that clearer and more accessible terms are beneficial for people who do read them. Clear and accessible terms can increase user understanding, which can in turn encourage rule following,[626] increase perceptions of platform fairness[627] and expand usership[628] among adults, indicating the same may be true for children.

19.49 Our analysis suggests four characteristics are important when determining whether provisions are clear and accessible to children. Evidence for each of these characteristics is explored below.

Ensuring provisions are easy to find

19.50 Ofcom research found that among 16–24-year-olds who had previously needed to access terms of service or guidelines on social media, 1 in 10 were unable to find them on at least one occasion.[629] A report from 5Rights highlights that terms can frequently be hidden within layers of menu options or split across multiple documents, as well as being more difficult to find after users have agreed to them.[630]

---

[622] Obar, J.A. & Oeldorf-Hirsch, A. 2020. The biggest lie on the internet: ignoring the privacy policies and terms of service policies of social networking services, *Information, Communication & Society*, 23 (1). [accessed 16 April 2024]. Subsequent references are to this article throughout.

[623] Ofcom, 2024. Online Platform Terms and Conditions and Content Controls. Question 1: When you sign up for social media or video sharing platforms, which of the following usually applies to you? Please select one only. Note: 58% of 16–24-year-olds reported that they usually agreed to terms and conditions without trying to access or read them, decreasing to 52% among all respondents; Obar & Oeldorf-Hirsch, 2020. 74% of US participants skipped the privacy policy when joining a service.

[624] Doteveryone (Miller, C., Kitcher, H., Perera, K. & Abiola, A.), 2020. People, power and technology: the 2020 digital attitudes report. [accessed 16 April 2024]. 45% of participants (aged 18+) said they often signed up to services online without understanding the terms; Unicef (Hartung, P.), 2020. The children's rights-by-design standard for data use by tech companies. [accessed 16 April 2024].

[625] Ofcom, 2024. Online Platform Terms and Conditions and Content Controls. Question 23 (TOS_23): If you were unsure about posting something on a social media or video sharing platform (in case it wasn't allowed), where would you check first to see if you should post it or not? Please select one option only.

[626] Matias, J.N., 2019. Preventing harrassment and increasing group participation through social norms in 2,190 online science discussions, *PNAS*, 116 (20). [accessed 16 April 2024].

[627] Jhaver, S. Appling, D.S., Gilbert, E. & Bruckman, A., 2019. 'Did you suspect the post would be removed?' Understanding user reactions to content removals on Reddit, *Proceedings of the ACM on Human Computer Interaction*, 3. [accessed 16 April 2024].

[628] Fiesler, C., Jiang, J., McCann, J., Frye, K., Brubaker, J.R., 2018. Reddit rules! Characterizing an ecosystem of governance, *Proceedings of the Twelfth International AAAI Conference on Wed and Social Media*. [accessed 16 April 2024]. Found that more popular subreddits by Reddit's ranking appeared to have more structured rules systems, suggesting clear and formalised rules made these communities more accessible to newcomers.

[629] Ofcom, 2023. Platform Terms and Accessibility. Question 2: You previously said you have needed to access the terms of service, community guidelines or any other type of policy document ('terms') of any social media website or platform...Which, if any, of the following describe your experience in trying to find information from these services' terms? (Please select all that apply). Note: 10% of 16–24-year-olds said they were not able to find them on at least one occasion.

[630] 5Rights Tick to agree, 2021.

19.51 Respondents to our 2023 Protection of Children Call for Evidence,[631] as well as relevant guidance,[632] make clear that for terms and statements to be accessible to children, they must be prominent, visible, and easy to find.[633] This would allow children to easily access and repeatedly visit terms and statements if they needed to, for example to check how a service deals with different kinds of harmful content. This should help to reinforce children's understanding of their rights and responsibilities as service users.[634]

19.52 More specifically, Carnegie UK advocate for terms and statements to be visible to would-be users before they sign up to a service,[635] allowing children and the adults who care for them to make an informed decision about the appropriateness of the service for children. This is particularly important given recent Ofcom research finding that terms for some prominent video sharing platforms were not accessible to non-users of the sites.[636]

19.53 In line with our recommendations around User reporting and complaints (Section 18), being able to find terms and statements is key to them being accessible to children. This means that they need to be intuitive to find and easy to reach through a small number of steps.

Laying out and formatting provisions to aid children's comprehension

19.54 Much research has explored the effectiveness of different presentation techniques to support the understanding of terms and statements.[637] For example, the Behavioural Insights Team found that summarising key terms using either icons or a question-and-answer format increased customers' understanding of terms by over 30% compared to the control.[638] The available research in this area has been conducted using adult samples, so we must exercise caution in interpretation. However, there is nothing to suggest that the same conclusions would not also apply broadly to children.

19.55 More specific insights about children's understanding of terms and statements come from existing guidance, reports and responses to our 2023 CFE. When consulted by different groups, children have expressed dislike for long paragraphs of text,[639] a desire for terms to

---

[631] NCMEC response to 2023 Protection of Children Call for Evidence; 5Rights response to 2023 Protection of Children Call for Evidence; Samaritans response to 2023 Protection of Children Call for Evidence.
[632] 5Rights Tick to agree, 2021.
[633] Findability is also championed by the ICO in their Age Appropriate Design Code (2020), indicating consistency of our proposed approach with another regulator.
[634] See for example Kang, S.H.K., 2016. Spaced repetition promotes efficient and effective learning: policy implications for instruction, *Policy Insights from the Behavioural and Brain Sciences*, 3 (1). [accessed 16 April 2024]. This article highlights the role of repetition in learning.
[635] Carnegie UK, 2023. Model code: A reference model for regulatory or self regulatory approaches to harm reduction on social media. [accessed 16 April 2024]. Subsequent references are to this document throughout; Carnegie UK response to 2023 Protection of Children Call for Evidence.
[636] Ofcom regulating video-sharing platforms (VSPs), 2023. "Snapchat and TikTok did not allow users to view their Community Guidelines if they were accessing via the app without an account." (p.15).
[637] For example, The Behavioural Insights Team Best Practice Guide, 2019; European Commission (Elshout, M., Elsen, M., Leenheer, J., Loos, M. & Luzak, J.), 2016. Study on consumers' attitudes towards Terms and Conditions (T&Cs): final report. [accessed 16 April 2024]; Danish Competition and Consumer Authority, 2018. Improving the effectiveness of terms and conditions in online trade. [accessed 16 April 2024]; Gage Kelley, P., Bresee, J., Cranor, L.F. & Reeder, R.W., 2009. A "nutrition label" for privacy. [accessed 16 April 2024].
[638] The Behavioural Insights Team Best Practice Guide, 2019.
[639] Mental Health Foundation response to 2023 Protection of Children Call for Evidence.

be clear about expected user behaviour[640] and a preference for key information to be presented using bullet points, a clear font and child-friendly imagery.[641]

19.56 Further suggestions for presenting child-friendly terms and statements included making them concise,[642] breaking them into clear sections,[643] making headings and key terms prominent,[644] layering additional detail under short notices of key information,[645] and giving examples to illustrate complex points.[646] Many sources also noted the importance of presenting terms and statements in an engaging way,[647] using graphics or icons to support children's understanding.[648]

19.57 5Rights highlight that services should not assume an engaged adult will be available to help children navigate terms and statements,[649] making the adoption of child-friendly presentation techniques particularly important.

19.58 Ofcom research found that 14% of 16–24-year-old respondents reported having difficulty reading information online due to illegible text because of weak contrast in colour between text and background.[650] To support people with a visual impairment, the Web Content Accessibility Guidelines recommend a 4:5:1 colour contrast between body text and background,[651] while Save the Children suggest including alternative text for all images and icons presented in terms and statements.[652] Similarly, to ensure accessibility for children with dyslexia, the European Commission advise employing a yellow background for terms and statements.[653]

Using clear and simple language to explain provisions

19.59 Ofcom's research found that among respondents (including 16- and 17-year-olds) who had reported being unable to get the information they needed from terms and statements, 55% said this was because the language was confusing, or written in a way that was difficult to

---

[640] Anti-Bullying Alliance response to 2023 Protection of Children Call for Evidence.
[641] European Commission, 2021.
[642] 5Rights response to 2023 Protection of Children Call for Evidence; Molly Rose Foundation response to 2023 Protection of Children Call for Evidence; The Children's Commissioner for England, September 2017. Simplified social media terms and conditions for Facebook, Instagram, Snapchat, YouTube and WhatsApp. [accessed 16 April 2024]; Ofcom video-sharing platform guidance, 2021; Schneble, Favaretto, Elger & Shaw, 2021.
[643] 5Rights response to 2023 Protection of Children Call for Evidence; IEEE standard, 2021; 5Rights Tick to agree, 2021.
[644] 5Rights response to 2023 Protection of Children Call for Evidence; IEEE standard, 2021; Ofcom video-sharing platform guidance, 2021; 5Rights Tick to agree, 2021.
[645] ICO response to 2023 Protection of Children Call for Evidence; ICO Age appropriate design code, 2020.
[646] Save The Children, 2022. How to Write a Child-Friendly Document. Subsequent references are to this document throughout.
[647] Glitch response to 2023 Protection of Children Call for Evidence.
[648] 5Rights response to 2023 Protection of Children Call for Evidence; ICO response to 2023 Protection of Children Call for Evidence; ICO Age appropriate design code, 2020; Save The Children, 2022; IEEE standard, 2021; 5Rights Tick to agree, 2021.
[649] 5Rights response to 2023 Protection of Children Call for Evidence.
[650] Ofcom, 2023. Platform Terms and Accessibility. Question 6: Now thinking about your time spent more widely online (i.e. beyond finding or reading terms)... Have you ever had difficulty reading information because of any of the reasons below? (Please select all that apply).
[651] Web Accessibility Initiative, 2023. Understanding SC 1.4.3: Contrast (Minimum) (Level AA). [accessed 16 April 2024].
[652] Save The Children, 2022.
[653] European Commission, 2021.

understand.[654] Just one of the six large video sharing platforms recently analysed by Ofcom provided Terms of Service likely to be understandable without at least a high school[655] education.[656] Among 16-24-year-olds who chose to accept terms without reading them, 26% did so because they felt they wouldn't be able to understand them, 40% did so because they found terms overwhelming, and 70% did so because they thought terms would take too long to read.[657]

19.60 The Behavioural Insights Team found that simplifying a policy's estimated reading age from a university graduate's reading level to a 14-year old's reading level increased comprehension by 16.9% among adults educated to GCSE level or below.[658] This suggests that a lower reading age would be beneficial for children and adult users.

19.61 Similarly, the Children's Commissioner for England said that a group of under 18s reported better understanding of Instagram's terms and conditions when they were presented using shortened and simplified language.[659]

19.62 Existing published guidance on presenting information to children, as well as responses to our 2023 CFE, highlighted that child-friendly terms and statements should use clear and age-

[654] Ofcom, 2023. Platform Terms and Accessibility. Question 4: You previously said that on at least one occasion, you were able to find the terms but could not get the information you needed from them. Why were you not able to get the information you needed from the terms? (Please select all that apply). Note: 55% of those who responded were able to find the terms but could not get the information needed stated confusing language as a reason. This question was based on a small sample of just 110 respondents.
[655] Finishing high school in the US is roughly equivalent to finishing sixth form in the UK (approx age 18), but the calculation cited used the US education system as a reference point, so we have retained the US-based language for accuracy.
[656] Ofcom regulating video-sharing platforms (VSPs), 2023. "TikTok's Terms of Service had the highest reading ease score (55) and it was the only platform where the Terms of Service were likely to be understood by users without a high school or university education. However, the reading level required was still higher than the typical reading level of the youngest users permitted on the platform." The other VSPs analysed were Snapchat, BitChute, Twitch, Brand New Tube and OnlyFans. (p.14).
[657] Ofcom, 2024. Online Platform Terms & Conditions and Content Controls. Question 15: You say you tend to accept platform T&Cs without reading them when signing up. Why is this? Please select all that apply.
[658] The Behavioural Insights Team Best Practice Guide, 2019. The study tested simplifying the Terms and Conditions of a peer-to-peer room sharing platform with sentences and words which were shorter on average. By doing this, they reduced the policy's estimated reading age from a university graduate's reading level to a 14-year old's reading level. Note: GCSEs are a UK educational qualification, usually undertaken by 15- and 16-year-olds to complete their secondary education.
[659] The Children's Commissioner for England Growing Up Digital, 2017.

appropriate language;[660] avoid jargon;[661] define difficult terms;[662] and address the reader directly using a human tone.[663]

19.63 Many respondents highlighted the importance of providing terms in multiple languages,[664] including British Sign Language, Easy Read and large print,[665] to ensure accessibility for all children.

Ensuring provisions are compatible with assistive technology

19.64 Around 11% of children in the UK were recorded as having a disability in 2021/22.[666] Some children with a disability may require certain tools to make use of terms and statements. For example, children with visual or motor impairments may be dependent on using a keyboard to navigate apps and webpages,[667] while screen readers make content on a screen accessible for those who are unable to see it.[668]

19.65 Provisions in terms and statements may not always be accessible to young internet users who rely on assistive technology. Ofcom research found that 18% of 16–24-year-old respondents reported having had difficulty reading information online in general because the content was not keyboard navigable, or was difficult to navigate using a keyboard. The same proportion reported the same difficulty because the content was not compatible, or was difficult to use, with a screen reader or screen reading technology.[669]

---

[660] Carnegie UK Model Code, 2023; Save The Children, 2022; ICO Age appropriate design code, 2020; 5Rights Tick to agree, 2021; European Commission, 2021; Designing for Children's Rights, 2022. Design Principles: Version 2.0. [accessed 16 April 2024]; International Telecommunication Union, 2020. Guidelines for industry on child online protection. [accessed 16 April 2024]; Carnegie UK response to 2023 Protection of Children Call for Evidence; Glitch response to 2023 Protection of Children Call for Evidence; NCMEC response to 2023 Protection of Children Call for Evidence; ICO response to 2023 Protection of Children Call for Evidence; Girlguiding response to 2023 Protection of Children Call for Evidence; The App Association response to 2023 Protection of Children Call for Evidence; 5Rights response to 2023 Protection of Children Call for Evidence; Resolver, a Kroll business (formerly Crisp) response to 2023 Protection of Children Call for Evidence.

[661] Samaritans response to 2023 Protection of Children Call for Evidence; 5Rights response to 2023 Protection of Children Call for Evidence; Ofcom video-sharing platform guidance, 2021.

[662] 5Rights response to 2023 Protection of Children Call for Evidence; ParentZone response to 2023 Protection of Children Call for Evidence; Save The Children, 2022; Ofcom video-sharing platform guidance, 2021; European Commission, 2021.

[663] Save The Children, 2022; Ofcom video-sharing platform guidance, 2021; European Commission, 2021; Samaritans response to 2023 Protection of Children Call for Evidence; ParentZone response to 2023 Protection of Children Call for Evidence.

[664] Antisemitism Policy Trust response to 2023 Protection of Children Call for Evidence; NCMEC response to 2023 Protection of Children Call for Evidence; 5Rights response to 2023 Protection of Children Call for Evidence; Carnegie UK response to 2023 Protection of Children Call for Evidence.

[665] Refuge response to 2023 Protection of Children Call for Evidence; SWGfL response to 2023 Protection of Children Call for Evidence.

[666] House of Commons Library (Kirk-Wade, E.), 2023. UK disability statistics: prevalence and life experiences. [accessed 16 April 2024].

[667] Web Aim, 2022. Keyboard Accessibility. [accessed 16 April 2024].

[668] Royal National Institute of Blind people, 2023. Screen Reading Software. [accessed 16 April 2024]; Ofcom video-sharing platform guidance, 2021.

[669] Ofcom, 2023. Platform Terms and Accessibility. Question 6: Now thinking about your time spent more widely online (i.e., beyond finding or reading terms)… Have you ever had difficulty reading information because of any of the reasons below? (Please select all that apply).

19.66 Commonly, terms and statements can include links at the top or side of the page. For users with certain disabilities, being able to skip links avoids the obstacle of navigating them to access the provisions.[670]

19.67 Semantic elements (the tags used to indicate what type of text is on the page) in HTML, which is the standard markup language for webpages, can also help those using screen readers and keyboards to navigate through information presented.[671]

19.68 The UN Commission on the Rights of the Child holds that terms and statements should be accessible to children of all needs.[672] In their response to our 2023 CFE, 5Rights advocated compliance with the latest Web Content Accessibility Guidelines as a potential means for ensuring such accessibility.[673] The Guidelines encourage reading sequences to be programmatically determinable, which is important for those using assistive technologies, and keyboard accessible.[674]

Rights assessment

19.69 This measure recommends that services likely to be accessed by children compile and present certain provisions within their terms of service or publicly available statement in a clear and accessible way, including for child users. In proposing this measure we have recommended that service providers should have regard to the findability and usability of these provisions, as well as how they are laid out and formatted and the language used to describe them. Whilst we have provided this steer in relation to this measure, this measure allows services flexibility in how they should achieve this outcome and it is not intended to be prescriptive.

19.70 The reasoning on rights to freedom of expression, association rights and users' (both children and adults) privacy rights that applies in relation to Measure TS1 above applies equally to this proposed measure. Our provisional conclusion is that this measure would not constitute an interference with users' (both children and adults) or services' freedom of expression or association rights. In addition, we similarly consider that this proposed measure is likely to achieve significant benefits for users in aiding understanding of the information which is of particular relevance to their experience on the service. These benefits may have positive impacts on users' - particularly children's - rights to freedom of expression and association, and also their rights to privacy in that it should help them understand how a service operates to protect them from content that might be harmful to them, and protect their personal data as they use and gain access to a service.

Impacts on services

19.71 The costs associated with this measure depend on the length of the relevant provisions, as the extent of information to be included in these provisions may vary between services.

---

[670] University of Washington, Access Computing, 2023. What is a skip navigation link?. [accessed 16 April 2024].
[671] MDN web docs, 2023. HTML: A good basis for accessibility. [accessed 16 April 2024].
[672] United Nations Committee on the Rights of the Child, 2021. General comment No.25 (2021) on children's rights in relation to the digital environment. [accessed 16 April 2024]. This position is also held by the IEEE (2021) in one of their voluntary process standards for age-appropriate digital design.
[673] 5Rights response to the 2023 Protection of Children Call for Evidence.
[674] Web Accessibility Initiative, 2023. Web Content Accessibility Guidelines (WCAG) 2.1 W3C Recommendation 21 September 2023 [accessed 16 April 2024].

These costs also depend on how comprehensible services' existing terms and statements regarding the protection of children are since this will determine the extent to which they would need to be revised. We do not expect these costs to vary greatly with the size of a service, though it is possible that the provisions which larger, more complex or high-medium risk services need to include to comply with the Act are longer. For example, these services might be using more measures to protect children, which they will need to include in their terms or statements. Additionally, such services might currently be using proactive technology to comply with any of the children's safety duties or may have longer processes around handling or resolution of complaints, details of which also need to be included in the provisions. Overall, the costs associated with the changes required to comply with this measure are likely to represent a higher share of revenue for smaller services with smaller budgets.

19.72 Moreover, the proposed measure is consistent with the equivalent measure presented in our Illegal Harms Consultation. We assume that services would follow both measures and therefore anticipate some synergies between the implementation of the two measures. We expect this to limit the additional costs associated with the measure we are proposing here. We also provide services flexibility on how they choose to apply the requirements set out above which allows them to tailor their approach to what is the most proportionate for them.

19.73 We estimate that services would need to incur costs between £3,000 and £5,000 to implement this measure. These costs represent the amount we expect services to incur to make terms and statements regarding the protection of children from harmful content clear and accessible. This is in addition to the costs incurred for implementing the corresponding Illegal Harms measure, which recommends that terms and statements regarding the protection of individuals from illegal content are made clear and accessible. Below, we present the breakdown of this overall cost estimate for each of the four characteristics proposed by the measure. The detailed assumptions underlying our cost estimates are found in Annex 12.

19.74 Our analysis suggests that the measure we are recommending will be effective in improving the clarity and accessibility of the provisions services will need to include in terms and statements. We consider that the costs for services in applying this recommendation will be relatively small and proportionate given the benefits to children and the adults who care for them in being able to access and understand important information about a service.

Ensuring provisions are easy to find

19.75 Services will need to ensure that the provisions presented in Measure TS1 are publicly available and easy to find. This would incur a one-off design and engineering cost to make the required user interface changes to meet this requirement.

19.76 In our Illegal Harms Consultation, we estimated that for most services, the one-off research and implementation cost of making provisions related to illegal harms findable would be between £2,000 and £5,000 and potentially significantly less for simple services that do not have a lot of functionality.[675] We also assumed there to be some smaller ongoing

[675] These figures assume that it would take up to five working days for a relevant employee to research the best ways to meet the requirements (assuming their salary is similar to a Software Engineer) and up to five working days for a Software Engineer to implement the changes. We consider these estimates to be at the higher end of the range as for many services it will take less time to research and implement any changes. This figure is based on 2022 prices and uses 2022 wage data.

maintenance costs. Since the required user interface changes would already have been made to comply with the equivalent Illegal Harms measure, we do not expect services to incur any additional costs over and above costs estimated in the Illegal Harms Consultation.

Laying out and formatting provisions to aid children's comprehension

19.77 Services will need to ensure that the provisions presented in Measure TS1 are laid out and formatted in a way that facilitates understanding among children, such as breaking text into segments and adding bullet points, child-friendly imagery, and prominent subheadings. Services may also need to decide on a text format, size, and colour relative to the background so that the text is easy to read. The cost impact of this is mitigated through services retaining flexibility on how they choose to help users, including children, read and understand their terms or statement, without the proposed measure making specific requirements.

19.78 The total cost would depend on the extent of revisions required by services and the specific choices made to achieve the outcome. These will largely be one-off costs, though services would also need to ensure they maintain suitable layout and formatting whenever they revise the provisions.

19.79 In our Illegal Harms Consultation, we anticipated the one-off research and implementation cost of ensuring provisions related to illegal harms are suitably laid out and formatted would be between £2,000 and £5,000,[676] with some smaller ongoing maintenance costs. To ensure that provisions regarding the protection of children are also formatted in a manner that facilitates understanding, including for children, we anticipate three working days of a software engineer's time for which we expect services to incur costs between £1,000 and £1,500, in addition to the costs estimated in the Illegal Harms consultation.

Using clear and simple language to explain provisions

19.80 Services may need to invest time and effort to ensure that the provisions presented in Measure TS1 are expressed in language that is comprehensible to the youngest person permitted to agree to them. The time and effort required will vary depending on the complexity of existing language used by services prior to implementing the current proposed measure as well as the youngest age that a service provider permits people to use the service without consent from a parent or guardian. In other words, the cost depends on the extent to which the provisions need to be revised. For example, the cost will be greater if there is a large difference in the reading age required to understand provisions and the age of the youngest person permitted to use the service without parental/guardian consent.

19.81 While making these changes would incur a one-off research and implementation cost, services will need to ensure that the same accessible language is used whenever they update these provisions. As an example, our Illegal Harms Consultation estimated that to simplify 800 words of text from a reading age of 16 to a reading age of 13, it would take a relevant employee three working days, costing the service between £500 and £1,500.[677]

19.82 We estimate that the current proposed measure may require higher costs. We anticipate that research and implementation would take eight working days of a relevant employee's

[676] This figure is based on 2022 prices and uses 2022 wage data, while the estimated figures for the current measure 2 (TS2) are based on 2023 prices and use 2023 wage data.

[677] This figure is based on 2022 prices and uses 2022 wage data, while the estimated figures for the current measure 2 (TS2) are based on 2023 prices and use 2023 wage data.

time for which we expect services to incur between £2,000 and £3,500 in addition to the costs estimated in our Illegal Harms Consultation.

Ensuring provisions are compatible with assistive technology

19.83 The provisions presented in Measure TS1 will need to be keyboard navigable and compatible with screen reading tools.

19.84 In our Illegal Harms Consultation, we anticipated the one-off research and implementation costs of making provisions related to the protection of individuals from illegal content usable would be between £2,000 and £5,000,[678] with some smaller ongoing maintenance costs. Since the required changes to make provisions keyboard navigable and compatible with screen reading tools would already have been made to comply with the equivalent Illegal Harms measure, we do not expect services to incur any additional costs beyond the costs estimated in our Illegal Harms Consultation.

Which providers we propose should implement this measure

19.85 We have provisionally concluded this measure should apply to providers of all U2U and search services likely to be accessed by children, as the Act requires these services to ensure that the provisions in their terms or statements outlined in Measure TS1 are clear and accessible.

19.86 We proposed an equivalent measure in our Illegal Harms Consultation. Therefore, assuming services would follow both measures, we expect services to save some costs in implementing the measure proposed here. We also provide services flexibility on how they choose to apply our requirements, which means they can tailor their approach to what is most feasible for them. Considering this and given the expected benefits of this measure in improving children's comprehension around keeping themselves safe on a service, we believe that this measure is proportionate.

Other options considered

19.87 We considered taking a prescriptive approach to this measure, recommending that services implement specific design criteria to achieve key characteristics of clear and accessible provisions for children. However, given the diversity and complexity of the services in scope of this measure, including their user bases and the design of their services, we do not consider that a prescriptive approach offers enough flexibility to achieve clarity and accessibility of the relevant provisions across these services.

19.88 Instead, we recommend that service providers achieve *outcomes* in line with the four characteristics of clear and accessible terms and statements set out above. We are confident that this will make our broad expectations clear to service providers, while allowing them more flexibility in the steps that could be taken to create clear and accessible provisions.

Provisional conclusion

19.89 This measure seeks to mitigate the risk of children and the adults who care for them not understanding children's rights and responsibilities as users of a service. We consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU D3 in Annex A7 and PCS D3 in Annex A8.

---

[678] This figure is based on 2022 prices and uses 2022 wage data.

Measure TS3 (Children's Safety Codes): Terms and statements for Category 1 and 2A services contain the findings of their most recent children's risk assessment

Explanation of the measure

19.90 In delivering this measure, we would expect to see providers of Category 1 and Category 2A services that are likely to be accessed by children develop or revise their terms or statement, ensuring they summarise the findings of their most recent children's risk assessment.

19.91 This measure is proposed for inclusion within our draft Children's Safety Codes. We are also proposing an equivalent measure for inclusion in our draft Illegal Content Codes (see below). This is in response to the recent publication of our categorisation advice to the Secretary of State in March 2024.[679] Providers in scope of both measures may operate a single version of their terms or statement to cover both measures if they wish.

19.92 The Act mandates that in scope Category 1 service providers and Category 2A service providers must summarise in their terms or statement the findings of their most recent children's risk assessment, including the levels of risk, and the nature and severity of potential harm to children.[680]

Rights assessment

19.93 This measure recommends that providers of in scope Category 1 and 2A services should summarise the findings of their most recent children's risk assessment in their terms or statement.

19.94 We consider that the reasoning set out in relation to Measures TS1 and TS2 above on the potential rights impacts of those proposed measures applies equally to this proposed measure. Our provisional conclusion is that this measure would not constitute an interference with users' (both children and adults) or services' freedom of expression, or association rights or users' privacy rights. Summarising levels of risk as well as the nature and severity of potential harm to children may result in positive benefits to users' - particularly children's - rights to freedom of expression and association, and also rights to privacy, in that it should also help them to understand why a service may elect or be required to implement particular measures in order to protect children from encountering children or contacts that might be harmful to them, as they use the service to express themselves and connect with other users.

Impacts on services

19.95 In scope Category 1 and 2A service providers who do not currently include provisions in their service's terms or statement that meet the relevant duty outlined above will need to add these provisions and incur the relevant costs. Since this measure reflects a direct requirement of the Act, any costs or impacts to services associated with this measure result directly from the duty in the Act. We have therefore not considered any costs or impacts to

[679] Ofcom categorisation advice, 2024.
[680] Sections 12(14) and 29(9) of the Act for full details of these duties.

services associated with this measure as part of assessing the implications of this measure for services.

19.96 We consider the requirements set out in the duty above are sufficiently clear for services to implement without further elaboration by Ofcom. We recognise that all service providers in scope of this measure would also be in scope of the equivalent measure being proposed for inclusion in our draft Illegal Content Codes (see below). While we expect services to incur incremental costs to meet the requirements of the current measure over and above the corresponding Illegal Harms measure, we have not considered these costs since the measure reflects a direct requirement of the Act.

Which providers we propose should implement this measure

19.97 This measure will apply to all Category 1 and Category 2A services that are likely to be accessed by children, as the Act requires these services to include in their terms or statements the relevant provision mentioned above.

Provisional conclusion

19.98 This measure seeks to mitigate the risk of children and the adults who care for them not understanding the risks posed to children as users of a service. We consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU D2 in Annex A7 and PCS D2 in Annex A8.

New Measure 6AA (Illegal Content Code): Terms and statements for Category 1 and 2A services contain the findings of their most recent illegal content risk assessment

Explanation of the measure

19.99 In delivering this measure, we would expect to see providers of all Category 1 and Category 2A services develop or revise their terms or statement, ensuring they summarise the findings of their most recent illegal content assessment.

19.100 This measure is proposed as a new addition to our draft Illegal Content Codes, and mirrors an equivalent measure proposed for inclusion in our draft Children's Safety Codes (see above). This measure was not previously included as part of our Illegal Harms Consultation because the consultation was published before we published our categorisation advice to the Secretary of State in March 2024.[681] Providers in scope of this measure and the equivalent protection of children measure may operate a single version of their terms or statement to cover both measures if they wish.

19.101 The Act mandates that all Category 1 service providers and Category 2A service providers must summarise in their terms or statement the findings of their most recent illegal content

[681] Ofcom categorisation advice, 2024.

risk assessment, including the levels of risk, and the nature and severity of potential harm to individuals.[682]

Rights assessment

19.102 This measure recommends that providers of all Category 1 and 2A services should summarise the findings of their most recent illegal content risk assessment in their terms or statement.

19.103 We have carefully considered whether this proposed measure would have any implications for freedom of expression or privacy for users (both children and adults). Our provisional conclusion is that it would not. This measure is intended to capture the specific requirements for in scope Category 1 and 2A services in relation to terms of service and publicly available statements under the Act. This does not require a service to take specific action in relation to content or personal information. We additionally consider that the provision of the specific type of information mandated by the Act and set out above, would be beneficial to users in that they would be consistently provided with information about the level of risk a service might pose to individuals.

Impacts on services

19.104 Category 1 and 2A service providers who do not currently include provisions in their service's terms or statement that meet the relevant duty outlined above will need to add these provisions and incur the relevant costs. Since this measure reflects a direct requirement of the Act, any costs or impacts to services associated with this measure result directly from the duty in the Act. We have therefore not considered any costs or impacts to services associated with this measure as part of assessing the implications of this measure for services.

19.105 We consider the requirements set out in the duty above, as well as the detail provided under Measure TS3, are sufficiently clear for services to implement without further elaboration by Ofcom. We recognise that all service providers in scope of this measure would also be in scope of the equivalent measure being proposed for inclusion in the Children's Safety Code (see above). While we expect services to incur incremental costs to meet the requirements of the current measure over and above the corresponding protection of Children measure (TS3), we have not considered these costs since the measure states a direct requirement of the Act.

Which providers we propose should implement this measure

19.106 This measure will apply to all Category 1 and Category 2A services, as the Act requires these services to include in their terms or statements the relevant provision mentioned above.

Provisional conclusion

19.107 This measure seeks to mitigate the risk of individuals not understanding the risks posed to them as users of a service. We consider this measure appropriate and proportionate to recommend for inclusion in the draft Illegal Content Codes. For the draft legal text for this measure, please see 6AA in Annex A9.

[682] See sections 10(9) and 27(9) of the Act for full details of these duties.

20. Recommender Systems

Recommender systems are a primary mechanism through which user-generated content is disseminated across U2U services. They are designed to make the service more appealing to users, matching them to content that is likely to be of interest, which results in higher content engagement and, often, an increase in the amount of time users spend on a service. For many service providers, recommender systems are essential for providing users with a selection of appealing and relevant content from the vast amount of content uploaded to their service.

However, evidence shows that these systems can also be a key pathway for children to encounter Primary Priority Content (PPC), including suicide, self-harm and eating disorder content, as well as pornographic content. They can also contribute to the amplification of other types of content that is harmful to children, for example violent content and content promoting abuse and hate. Additionally, recommender systems play a part in narrowing down the type of content presented to the user, which can lead to increasingly harmful content recommendations ('rabbit holes') as well as the risk of cumulative harm.

What is cumulative harm and what difference will our proposals for recommender systems make?

As described in the draft Children's Register of Risks (section 7), cumulative harm can occur when content that is harmful to children is repeatedly encountered and/or when a children encounter harmful combinations of content. This can also occur when children encounter content that is harmful to them (as defined by the Online Safety Act 2023 ('the Act')) alongside content that could be harmless. For example, dieting content in and of itself may not be harmful but when encountered alongside content that promotes eating disorders (PPC) this could be extremely harmful, and result in cumulative harm.

Our proposals recommend that U2U services operating a recommender system, and posing a risk of exposing children to content harmful to them, follow a precautionary approach to content shown in children's feed. This is achieved through filtering out content likely to be PPC (Measure RS1) and limiting the prominence of content likely to be PC (Measure RS2). On large risky services, children should also be offered more control, allowing them to indicate if they do not want to continue to see certain types of content (Measure RS3).

We have assessed the potential impacts, including costs and rights impacts, of these proposals and consider them to be proportionate for the services suggested to be in scope of these measures, given the risks that recommender systems pose to children.

Our proposals

#	Proposed measure	Who should implement this[683]
RS1	Ensure **that content likely to be PPC is not recommended to children.**	All U2U services that • Have content recommender systems; and • Are medium or high risk for at least one kind of PPC
RS2	Ensure that **content likely to be PC* is reduced in prominence on children's recommender feeds**	All U2U services that • Have content recommender systems; and • Are medium or high risk for at least one kind of PC (excluding bullying)
RS3	**Enable children to provide negative feedback** on content that is recommended to them	All U2U services that • Have content recommender systems; and • Are medium or high risk for at least two kinds of PPC and/ or PC (excluding bullying)**; and • Are large

* We are also minded to include two potential kinds of NDC, subject to consultation. If we do recommend that these kinds of content are classified as NDC, then RS2 would be recommended for all U2U services that have content recommender systems and are medium or high risk for at least one kind of PC (excluding bullying), body image, or depressive content;
** If we do recommend these kinds of NDC, RS3 would be recommended for all large U2U services that have content recommender systems and are medium or high risk for at least two kinds of PPC, PC (excluding bullying), body image, or depressive content

Consultation questions

49. Do you agree with the proposed recommender systems measures to be included in the Children's Safety Codes? Please confirm which proposed measure your views relate to and provide any arguments and supporting evidence. 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.

50. Are there any intervention points in the design of recommender systems that we have not considered here that could effectively prevent children from being recommended primary priority content and protect children from encountering priority and non-designated content?

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.

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.

[683] These proposed measures relate to providers of services likely to be accessed by children.

What are recommender systems and how do they work?

20.1 **Recommender systems** are comprised of algorithms that learn about users' interests based on factors such as user behaviour and personal characteristics to select content that may be of interest to them. Based on this information, algorithms enable recommender systems to tailor content to specific users. Recommended content is sourced widely and often comes from accounts that the user has not chosen to connect with or follow.

An explainer: What is a recommender system?[684]

Recommender systems provide personalised content recommendations to users based on information about their personal characteristics (e.g. age, location, gender) and historic engagement on the service (such as watch time, likes, reshares). Recommender systems that curate feeds of content (such as newsfeeds and reels) for users on U2U services are known as **content recommender systems.** These systems are powered by algorithms.

An **algorithm** is a sequence of computational instructions that help a programme or application achieve a specific goal.[685] Content recommender systems use different kinds of algorithms to learn about content types, user preferences, and match users to content. In addition to personalisation, content recommender systems can be designed to offer content variety, taking into account the diversity and popularity of content on a service.[686]

The measures proposed in this chapter would only apply to **content recommender systems**, and not to those systems that underpin search functionalities on a U2U service, or network recommender systems that suggest other users to follow or groups to join. For the remainder of this chapter the term recommender systems will refer to content recommender systems.

20.2 **Recommender systems** are made up of many different algorithms, which rely on machine learning. Algorithms help the recommender system by identifying relationships between content and users by analysing content features and characteristics, as well as learning about users' content preferences and their characteristics. These algorithms include, but are not limited to, **scoring algorithms** and **re-ranking algorithms**.

20.3 **Scoring algorithms** predict what content the user is most likely to engage with. Effectively, they give content a predicted engagement "score" for each user, which represents the likelihood that the user will engage with the content (such as watch it, share it, like it, or comment on it) – this is also known as the relevancy score. Based on the relevancy score, content will be ranked accordingly. This is known as engagement-based ranking.

20.4 One of the ways that scoring algorithms learn about user preferences and curate content for them is through **collaborative filtering**. This means that users that have similar engagement patterns (for instance, following the same pages and watching similar content) will mutually influence one another's content recommendations. If person A and person B have a similar taste in a particular type of content, the recommender system is designed to infer that these

[684] Definitions of key terms including recommender systems and content recommender systems can also be found in the Glossary (Annex 15).
[685] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.
[686] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.

users may have the same taste in other types of content. This can often result in user clustering, where content is scored similarly for users that are regarded as having shared characteristics. If a child explicitly engages with that content by liking it or sharing it, they are likely to be sending positive signals to the recommender system. Depending on the service's design choices, children may also implicitly "engage" with content simply by clicking on it or hovering over it, even though the engagement could be drawn from negative feelings, such as shock or disgust[687] rather than enjoying the content.

20.5 Content recommender systems often use **re-ranking algorithms** to refine the initial list of recommendations curated by scoring algorithms.[688] During the re-ranking stage, content can be reordered to ensure users are presented with more diverse and novel content. This stage is also often the point where services may remove or limit the prominence of harmful content. At the re-ranking stage, we understand recommender systems can be designed to respond to a variety of **relevant available information** from users and other processes for the purpose of applying safety measures.[689]

20.6 As described below in Box 1, content moderation is one of the processes that provides recommender systems with relevant available information.[690] Content moderation provides signals about what content should be removed from the service or should be limited in visibility. Based on our understanding of the systems, we expect services that have recommender systems, also have the capabilities to classify and categorise content at scale, to inform and optimise recommendations for users, often by means of automated content classifiers.

20.7 While the use of automated systems is of value for the deployment of the measures outlined in this chapter, alongside information gained as part of moderation processes, services also have access to a range of **relevant available information** that they can use to inform what content is disseminated to children by means of recommender systems.

Definition Box 1: What relevant available information might recommender systems be able to use to indicate that content is likely to be harmful to children?

Using their existing systems and processes, there are a number of ways that services can become aware that content may be harmful to children. Relevant available information means any kind of information or signal that can act as an indicator for the recommender system to determine the appropriateness of content for recommendation to children. Relevant available information that can act as a signal to the recommender can be generated in a variety of ways including, but not limited to:

Content identification processes: automated content classifiers (e.g., machine learning and heuristic techniques) and trained moderators can assess whether content is likely to be harmful to children or not and can label content. For example, content identified as likely to be harmful might be labelled as 'violent' meaning that the algorithm can filter this out so that it is not

[687] Ofcom, 2022. Research into risk factors that may lead children to harm online; Integrity Institute, 2024. Child Safety Online. 19 January. [accessed 19 January 2024].
[688] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content
[689] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.
[690] Service providers should also have regard to the ICO's guidance relating to content moderation and data protection.

recommended to children.[691] Where content has completed the moderation process and has been found to not be harmful to children, this may re-enter the recommender systems for children.

User feedback, reports, or tags at upload: Services can collect feedback from users about content and use this information to inform the recommender system. Examples include user complaints or reports.[692] Services may also have feedback from trusted flaggers and other information such as ratings applied by users at upload which may indicate that content is likely to be harmful to children.

Information available due to other codes measures: For example, services that implement Measure RS3 (providing children with a means of expressing negative sentiment) will have negative feedback signals from children that should be considered relevant available information for Measure RS1 and Measure RS2 described in this chapter.

Beyond those listed here, services may have other kinds of relevant available information such as an indication that accounts are highly likely to have content that is harmful to children based on the description of the user account and the kinds of content shared. Recommender systems should be designed to take a **precautionary approach** and use this information as instructions to filter out and manage the volume and prominence of potentially harmful content shown to children as set out in these measures.

What risks do recommender systems pose to children?

20.8 Recommender systems are designed to make the service more appealing to users, matching them to content that is likely to be of interest, which results in higher content engagement maximising the time a user spends on the service and the service's profitability.[693] Recommender systems can, however, also be designed to take into account the safety of children, by ensuring that the content served by the recommender system is not harmful. For more detail on the risks associated with content recommender systems, refer to Section 7.14, Volume 3 focused on the Wider context to understanding risk factors.

20.9 The algorithms that make up recommender systems often rank content based on the likelihood that users will engage with it. The use of such 'engagement-based design' can risk exposing children to more harmful content [694] [695] by introducing children to such content for the first time. Engagement based ranking can perpetuate users' vulnerabilities by using harmful content engagement habits as positive signals to recommend more of the same

[691] Thorburn, L, Bengani, P, Stray, J., 2022. How platform recommenders work. Medium, January 20 2022. [accessed 11 April 2024].
[692] Measure UR1 recommends that services have complaints processes which enable users to make complaints.
[693] 5Rights Foundation, 2021.Pathways: How digital design puts children at risk. [accessed 11 April 2024].. Note: The research involved setting up a series of avatars, which were profiles set up on social media apps that mimicked the online profiles of real children who took part in the interviews for this project. The age of the real child was used to register the profile and displayed in the bio of the user account. [accessed 11 April 2024].
[694] Integrity Institute, 2024. Child Safety Online. [accessed 19 January 2024].
[695] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.

type of content.[696] For example, a child seeking out eating disorder content is likely to be particularly vulnerable to harm from this kind of content, yet current service design means that more vulnerable children are more likely to be serviced high volumes of eating disorder content, leading to cumulative harm.[697]

20.10 We also have evidence that children may reluctantly engage with or watch content recommended to them because of the perceived popularity of this content among peers; children in the research said they felt they had no control over the content recommender systems suggested, and therefore seeing more violence on their feed felt inevitable.[698] There is also evidence that users (including children) liking and/or resharing content also provides positive user feedback that can feed into the virality of online content.[699] For example, the liking and re-sharing of content depicting dangerous stunts and challenges can increase the likelihood of children encountering this content.[700] Children may also engage with content even though it is harmful to them, or they initially react with shock or disgust causing them to hover over content. For instance, a survey with 11–16-year-olds found that on first viewing pornography, children often reported feeling shocked or confused.[701] However after repeated exposure, feelings of shock and confusion dissipated as they became seemingly desensitised to the content.[702]

20.11 This is particularly the case when recommender systems are designed to interpret implicit engagement (such as hovering on content) as user preferences. How a content recommender system is designed can therefore influence the extent to which certain categories of PPC and PC are disseminated on a service and, in turn, increase the risks posed to children. Refer to the draft Children's Register of Risks for more detail.[703]

20.12 We understand that, even though many services already prohibit some content that is harmful to children under the Act in their terms of service, this content is often available on services and accounts registered as children are able to access this. This risks content that is harmful being presented to children and amplified by the recommender system.[704] Despite a

---

[696] Corrected oral evidence: Consideration of government's draft Online Safety Bill, Monday 25 October 2021. Q163, page 12. [accessed 14 December 2023].

[697] See the draft Children's Register of Risks (section 7), in particular sub-section 7.3 relating to eating disorder content.

[698] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.

[699] Ofcom, 2023. Evaluating recommender systems in relation to the dissemination of illegal and harmful content in the UK.

[700] See the draft Children's Register of Risks (section 7), in particular sub-section 7.8 relating to dangerous challenge content.

[701] Martellozzo, E., Monaghan, A., Adler, J.R., Davidson, J., Levya, R. and Hovarth, M.A.H., 2017. 'I wasn't sure it was normal to watch it'. [accessed 20th June 2023]

[702] Martellozzo, E., Monaghan, A., Adler, J.R., Davidson, J., Levya, R. and Hovarth, M.A.H., 2017. 'I wasn't sure it was normal to watch it'. [accessed 20th June 2023]

[703] See the draft Children's Register of Risks (section 7), in particular sub-sections related to pornography, eating disorder content, suicide and self-harm content and abuse and hate, violent content, dangerous stunts and challenges and harmful substances.

[704] The Bright Initiative and Molly Rose Foundation, 2023. Preventable yet pervasive: The prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest. Note: In this study the researchers explored Instagram, TikTok, and Pinterest with avatar accounts registered as being 15-years-of-age. Content was identified and scraped using hashtags that have been frequently used to post suicide and self-harm related material. While this is a singular study and may not represent all children's experiences, it demonstrates that this type of content was available on the services at the time of the study. [accessed 27 March 2023].

range of relevant available information which can serve as signals to the recommender system (see Box 1), content that is harmful to children is not currently consistently filtered out or reduced in prominence in the recommender feeds for children.

20.13 Our research shows that there is widespread availability of content that is harmful to children and that recommender systems are a key pathway for children to encounter PPC and PC. This can include introducing children to this content for the first time. For example, there is evidence that there are children who have encountered PPC online, in particular, suicide, self-harm and eating disorder, without seeking this out, with children feeling that this content had been shown to them because of the service's recommender system.[705] As described in the draft Children's Register of Risks, recommender systems may also lead children to encounter PC, including abusive content, content inciting hatred, and content depicting or encouraging violence. There is also some evidence that dangerous online stunts and challenges can become viral via recommender systems.[706]

20.14 Research by the 5Rights Foundation found that the U2U services in scope of their research typically designed their recommender systems to prioritise user engagement. This research also found that the accounts used in the research that were registered as children were being targeted with age-specific advertising but were being recommended content that was not appropriate for their age and was often harmful to children.[707] This suggests that signals and predictions about user age or inferred interests can be used to recommend specific content (e.g., advertising content), but these relevant signals are not always leveraged to consistently filter out content that may be harmful to children.

20.15 We know that children can find encountering harmful content distressing, particularly when this is unintentional. According to the NSPCC, some young people who spoke in Childline counselling sessions about viewing harmful content had come across this material unintentionally while browsing online spaces which they believed to be safe, making them feel unnerved and uncomfortable.[708]

20.16 The evidence shows that children are at risk of being introduced to harmful content where a recommender system uses collaborative filtering algorithms (See above sub-section 'What are recommender systems and how do they work?') on a service that may not effectively distinguish between adult and child users. In this scenario, there is an increased risk of children being grouped with adult users who might be engaging with, for example, suicide and self-harm content. This risk can also occur between different children where some child users are engaging with harmful content. Since collaborative filtering algorithms use shared interest as the basis for inferring interest, this may result in harmful content being introduced to a child by association, even if they were not intentionally looking for it.[709] For example, research by 5Rights included examples of some platforms recommending harmful

---

[705] Ofcom, 2024. Online Content: Qualitative Research, Experiences of children encountering online content promoting eating disorders, self-harm and suicide.

[706] See the draft Children's Register of Risks (section 7), in particular sub-sections related to abuse and hate content (7.4), and violent content (7.6). See also sub-sections relating to dangerous stunts and challenges content (7.8).

[707] 5Rights Foundation, 2021. Pathways: How digital design puts children at risk. Note: The research involved setting up a series of avatars, which were profiles set up on social media apps that mimicked the online profiles of real children who took part in the interviews for this project. The age of the real child was used to register the profile and displayed in the bio of the user account. [accessed 11 April 2024].

[708] NSPCC, 2022 Children's experiences of legal but harmful content online. [accessed 1 March 2024]

[709] 5Rights Foundation, 2021. Pathways: How digital design puts children at risk. [accessed 11 April 2024].

content to accounts registered as children. The platforms cited are ones where it was likely collaborative filtering was being used.

20.17 Further, our understanding is that all recommender systems are likely to play a role in causing harm to children through gradual exposure to increasingly harmful content. This can occur when the recommender system responds to users who may be engaging with positive or neutral content by offering more extreme content that may be deemed more engaging. For example, there is evidence to suggest that engaging with content relating to weight loss online, which could be diet content, can lead users to encounter content promoting eating disorders.[710] We also understand that some users have been recommended eating disorder content after engaging with recovery or support content.[711] Similarly, we have evidence that children who encounter violent content may become desensitised to this content and that this can mean children do not consider reporting this,[712] which could risk children seeing more of this harmful content.

20.18 Repeated engagement with a particular type of content can result in a **"**filter bubble**"**[713] whereby a user's feed is increasingly filled with a specific type of content, and they are recommended fewer alternative types of content, narrowing their field of interest. If the content is harmful, this can result in the user being presented with an increasing volume of harmful content, often increasingly harmful and extreme. This is what is also known as the **"**rabbit hole" effect.[714]

20.19 We also know that repeatedly viewing PPC on recommended feeds, in particular suicide, self-harm, and eating disorder content, as well as content related to this such as what we are proposing may be identified as 'depressive content' and 'body image content' (potential NDC, subject to consultation (see Section 7.9, Volume 3) can result in children experiencing cumulative harm. Recent research by the Molly Rose Foundation has highlighted the potential risk of cumulative harm posed by certain types of content viewed in large amounts. The report describes how this poses the most substantial risk to "children and young people experiencing suicide ideation, thoughts of self-harm or poor mental health."[715] Similarly, our own risk factors study showed that cumulative passive exposure to hazards over time can build up to cause more significant harm.[716] In their response to our 2023 Call for Evidence (our 2023 CFE), the Molly Rose Foundation described how the thousands of

---

[710] 5Rights Foundation, 2021. Pathways: How digital design puts children at risk. [accessed 19 April 2024]; Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content

[711] Beat response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

[712] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.

[713] The term 'filter bubble' was coined by Eli Pariser in his 2011 book *The Filter Bubble: What the Internet is Hiding from You.* Source: UK Parliament, House of Commons Library, 15 January 2024, Preventing misinformation and disinformation in online filter bubbles. [accessed 21 April 2024].

[714] A 'Rabbit hole' is the process of recommending ever more extreme content to users over time, which may occur as a result of users engaging with that type of content in the past. Particularly likely among users who already exist in filter bubbles. Source: Ofcom, 2023. Evaluating recommender systems in relation to the dissemination of illegal and harmful content in the UK.

[715] The Bright Initiative and Molly Rose Foundation, 2023. Preventable yet pervasive: The prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest. Note: In this study the researchers explored Instagram, TikTok, and Pinterest with avatar accounts registered as being 15-years-of-age. Content was identified and scraped using hashtags that have been frequently used to post suicide and self-harm related material. While this is a singular study and may not represent all children's experiences, it demonstrates that this type of content was available on the services at the time of the study. [accessed 27 March 2024].

[716] Ofcom, 2022. Risk factors that may lead children to harm online.

pieces of harmful content algorithmically recommended to Molly had a long-term and cumulative effect on her.[717]

20.20 Our detailed evidence around the risks posed by recommender systems is captured in Volume 3 detailing the causes and impacts of harm and the Governance and Accountability Measures in Section 11 of Volume 4.

Our proposals to protect children

20.21 The Act requires all providers of U2U services to take measures if it is proportionate to do so with regard to the "design of functionalities, algorithms and other features" in delivering the safety duties protecting children.[718]

20.22 In developing our proposals for how service providers can meet these duties, we consider that designing recommender systems with the safety of children as a priority is a key intervention service providers are able to make.[719]

20.23 Our three proposed measures focus on changes to the design of services' recommender systems to protect children from harm and deliver materially better outcomes for children:

- **Measure RS1:** We recommend that all U2U services likely to be accessed by children that are medium or high risk for any kind of PPC (regardless of size) **and** have content recommender system functionality, design their recommender systems to filter out content likely to be PPC from the recommender feeds of children.
- **Measure RS2:** We recommend that all U2U services likely to be accessed by children that are medium or high risk for any kind of PC (regardless of size) **and** have content recommender system functionality, design their recommender systems to reduce the prominence of content that is likely to be PC in the recommender feeds of children.
- **Measure RS3:** We recommend that all large[720] U2U services likely to be accessed by children that are medium or high risk for two or more kinds of PPC and/ or PC **and** have content recommender system functionality provide children with a means of expressing negative sentiment and feedback directly to the recommender feed, on content they encounter on recommender feeds.

20.24 We are currently consulting on potentially identifying two kinds of non-designated content, namely body image content and depressive content, subject to further evidence on defining these kinds of content and identifying a link to significant harm. See Volume 3, Section 7.9 (Non-designated content). If we are in a position to identify specific categories of NDC in our Children's Register of Risks in our final statement, we are minded to recommend that

[717] Molly Rose Foundation response to 2023 Protection of children Call for Evidence; The Bright Initiative and Molly Rose Foundation, 2023. Preventable yet pervasive: The prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest. [accessed 27 March 2024].

[718] Section 12(3)(a) and (b) and section 12(8)(b) and (f) of the Act.

[719] We have consulted with industry experts as part of our consideration of these measures. This included engagement with Rumman Chowdhury, Founder of Humane Intelligence and Ravi Iyer, Managing Director of the USC Marshall School's Neely Center. In addition to providing expertise on recommender systems more generally, this expert input helped to inform our consideration of the technical feasibility and costs associated with these measures.

[720] See Framework for Codes at Section 13 within this Volume for a definition of a large service.

Measures RS2 and RS3 also cover such non-designated content, where content recommender systems are a risk factor for this kind of content.

20.25 Services need to have the right systems and processes in place to ensure the effectiveness of this measure at reducing the risk of exposure of children to harmful content.

20.26 Services in scope of these measures must secure that all users who may be children (i.e., are not determined to be adults, which would include logged-out users who have not undergone any form of age assurance) have content likely to be PPC filtered out and PC significantly limited in their recommended feeds.

20.27 In targeting the measures to children, service providers are recommended to do so by means of highly effective age assurance to correctly identify child users. It is for providers to determine when to implement highly effective age assurance so long as the relevant recommender systems measures apply to all child users' recommender feeds whether logged in or out. Refer to Section 15 of this Volume for more information on the requirements and the role of highly effective age assurance in protecting children from harmful content.[721]

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

Explanation of the measure

20.28 In delivering this measure, we would expect service providers in scope of this measure to adjust the design of their recommender systems to consistently filter out content that is likely to be PPC and not include this in recommender feeds for children. The benefits of this proposed measure, in terms of improved protection of children, are partly contingent on Measures AA5 and AA6 set out in Section 15, Age Assurance, within this Volume.

20.29 The Age Assurance measures sets out that highly effective age assurance should be used by services in scope of this measure to apply this measure to children. This ensures that children are protected by this measure as opposed to applying to all users. Service providers in scope of this measure will incur costs associated with both the proposed measure requiring services to filter out PPC from the recommender system and the relevant proposed measure described in Section 15, Age Assurance within this Volume. While we discuss these measures separately in the respective sections, we have had regard to the combined costs and benefits in the round as part of our assessment and explanation of how the measure works.

20.30 To implement this measure, the service provider in scope of this measure should use existing relevant available information (see Definition Box 2 for some examples of this) to determine if content is likely to be PPC. We recognise that relevant available information may vary in accuracy and quality. For example, this could be content that has been reported by a user but not yet moderated. We are not prescriptive about when services consider that content is likely to be PPC for the purposes of this measure. For example, the number of

[721] Service providers should also have regard to the ICO's Opinion on Age assurance for the Children's code.

user reports needed to consider content likely to be PPC could be an existing threshold for determining when content is further moderated, or it could be a lower threshold.

20.31 While this threshold should be determined by the service provider, given the serious risk of harm to children, we are proposing that service providers in scope of this measure take a precautionary approach to ensure that the relevant instructions are sent to the recommender system. This is to ensure that any content that is likely to be PPC is not recommended to children, whether it has yet been confirmed as such or not through the content moderation processes. In addition, the flexibility given to services to set their own thresholds as to when content becomes likely to be PPC, for example, the number of complaints or reports received, should not have the effect that services ignore relevant available information. Where this measure is deployed effectively, services should consistently filter out content that is likely to be PPC from the recommender feeds for children.

20.32 We are taking a flexible approach to the relevant available information that service providers use to inform the above. All service providers are expected to have existing sources of relevant information. Service providers may choose to introduce additional ways to gather relevant information that content is likely to be PPC but are not required to. There may also be additional relevant available information if services implement other measures proposed in the code.

20.33 Removing this content from children's feeds will impact the unintentional exposure of children to this content, without impacting the reach of the content for adults and those children actively seeking content harmful to children as this may still be accessible on other parts of the service. Refer to the Content Moderation Measures for U2U services , Section 16 of this Volume for proposed measures for what services should do to identify and moderate content.

Definition Box 2: What is content *likely* to be PPC?

Primary Priority Content (PPC) includes: Pornographic content; content which encourages, promotes or provides instructions for suicide; content which encourages, promotes or provides instructions for an act of deliberate self-injury; and content which encourages, promotes or provides instructions for an eating disorder or behaviours associated with an eating disorder.[722]

More information about the definitions of PPC can be found in our Guidance on content harmful to children in Volume 3, Sections 8.2-8.5. This includes examples of content or kinds of content that we consider to be, or not to be, PPC.

For the purposes of this measure, content likely to be PPC is expected to include:

- content which is undergoing content moderation or has been flagged for moderation.
- content which has not undergone any form of content moderation, but where relevant available information indicates that there is a material likelihood of that content being PPC.[723]

[722] Section 61 of the Act.

[723] Separately, Measure CM4 in Section 16 of this Volume recommends that services have regard to whether content is likely to be PPC in deciding which moderation cases should be prioritised. RS1 goes further by requiring action to be taken on "likely to be" content in recommender systems, even though a piece of content does not yet have a final moderation determination.

U2U services should leverage their existing capabilities and utilise relevant available information as instructions for the recommender systems to filter out content likely to be PPC.

20.34 As set out in Definition Box 2 there are a number of ways a service may become aware of content that is *likely to be* PPC. For example, content moderation methods that services may employ to identify and label content that is likely to be PPC.[724] These often include automated content moderation (ACM) tools, often in tandem with human review to help confirm the nature of the content.[725] Services may also use user rating systems, which are tools that allow users (uploaders and viewers) to rate content.[726] This involves adding labels to content to denote that it may be unsuitable for certain audiences, such as children.

20.35 This proposed measure (RS1) does not specify any specific methods for how services can become aware of content that is likely to be PPC, but rather provides examples of the kinds of relevant available information that could be used to do so in Definition Box 1. This is also consistent with our approach in relation to Content Moderation Measures for U2U services outlined in Section 16 of this Volume which will enable flexibility in how services establish how to identify and moderate content. Where relevant available information exists under a service's existing capabilities, we expect it to be utilised for this measure. Service providers may also decide to further moderate content that is flagged as likely to be PPC to ensure that appropriate action is being taken to protect children from being exposed to PPC, in line with our recommended measure on content moderation.

20.36 While we are proposing to allow flexibility in the way services identify and label content, we are expecting services to take into account the wide range of relevant available information they have, and to apply a precautionary approach, when deciding what content should be served to children through the recommender system. Signals can include information that is not resulting from content moderation systems, for example tags applied by users at upload. Service providers may choose to combine sources of available relevant information to assess if content is likely to be PPC, for instance using an inference model to incorporate different sources. Service providers could also consider where there are inferences which the recommender system itself can make to support the identification of likely PPC. The flexibility of our measure allows a service provider to best use the information it has available to detect likely PPC, and leaves scope for innovative methods to detect such content.

20.37 When implemented effectively, this measure will consistently filter out content likely to be PPC for all children. This will reduce the likelihood of children being exposed to PPC when they are not looking for it, thereby preventing some children from encountering this content in the first place and will reduce the likelihood of a potential 'rabbit hole' effect which

[724] Many services take a hybrid approach to content moderation, i.e. using both human and automated resources.

[725] Grindr, for example, states that it uses 'proprietary technological tools' to help it proactively flag illicit content. Source: Grindr, no date. How Grindr moderates content and profiles . [accessed 19 April 2024]. Meta states it is increasingly using an 'automation-first approach' to content moderation to review more content across all types of policy violations. Source: Meta, 2020. How We Review Content. [accessed 19 April 2024]. In its response to the 2022 Illegal Harms Call for Evidence, Roblox told us it deploys 'several automated systems' that will 'scan files for illegal content and egregious violations' of Community Standards, in addition to the use of human moderators. While we know some services use various forms of automated content moderation (ACM) tools to identify content for moderation, we currently have limited information about most of these.

[726] For example, service providers such as Vimeo, Twitch, Tumblr use user rating systems.

involves recommended content becoming more harmful over time. It will also ensure that children who may be actively seeking harmful content will not be subsequently continuously pushed this type of content by the recommender system.

20.38 For this measure we provisionally recommend that services make any necessary adjustments to their recommender system to ensure that content that is likely to be PPC is filtered out before it is recommended to children. We consider that the steps services may need to take to do this are:

i) Use relevant available information to identify content likely to be PPC;
ii) Make the signal available to the recommender system; and
iii) Modify the recommender system to filter out content likely to be PPC from content recommended to children.

Effectiveness at addressing risks to children

20.39 We consider this measure to be, at a systems and processes level, an effective way for services to prevent children from encountering PPC. Research shows that suicide, self-harm, and eating disorder content is highly prevalent and available to children on some U2U services popular with children. Recommender systems heighten the risk of harm to children by driving exposure to harmful content that exists on those services. As detailed in the draft Children's Register of Risks in Volume 3, Section 7, recommender systems are a key pathway for children to encountering suicide, self-harm, and eating disorder content.

20.40 Our evidence shows children often encounter PPC on their recommended feeds unexpectedly or accidently, even where they have not previously searched for or interacted with it. More detail on this can be found in the Governance, systems and processes sub-Section 7.11, Volume 3. We consider the measure will reduce the likelihood of accidental or unexpected viewing by filtering out content that is likely to be PPC content on recommender feeds. This may also prevent children's initial exposure to this content, which our evidence shows often happens via recommender systems. Similarly, based on the evidence in Volume 3, Section 7.1, focused on pornographic content this measure would prevent children from being recommended pornography.

20.41 Children who have had some engagement with suicide, self-harm, or eating disorder content previously, may be more likely to be recommended more of this content.[727] This is likely to include those already at a heightened risk from this type of content or those who may be experiencing mental health difficulties or other vulnerabilities. Our evidence suggests that these groups are also more likely to proactively seek out content, including in isolated events of crisis, which could, in turn, risk these children being recommended more of it. We consider this measure could mitigate the risk that these children are then recommended more of this content after seeking it out, reducing the risk of continuous exposure and ongoing engagement.

20.42 As our evidence set out in the governance, systems and processes section (Volume 3, Section 7.11) of this consultation indicates, recommender systems can lead to a potential rabbit hole effect[728] where engagement with certain topics can, over time, lead to

[727] For more detail, please refer to the Governance, systems and processes sub-section (7.11) in Volume 3 in this consultation.
[728] For more detail please refer to the Governance, systems and processes sub-section (7.11) of Volume 3 in this consultation.

recommendations of content that is increasingly extreme or harmful in nature (e.g. healthy eating content may lead to restricted eating content). Where a child engages with content relating to a topic that is thematically adjacent to a harm, this measure will help minimise the risk of a child repeatedly (or increasingly) being recommended content that is harmful, such as content with themes adjacent to suicide, self-harm, and eating disorders[729] which could lead to recommendations of PPC.

20.43 Tragic cases, such as those of Molly Russell[730] illustrate the potential cumulative impact and risk of harm amounting from sustained exposure to suicide, self-harm and eating disorder content by means of recommender systems.[731] Although this measure may not prevent children from encountering PPC on all parts of a service, for example if searching for content on a U2U service, we consider that it would significantly reduce the cumulative harmful impact which arises from repeatedly encountering content likely to be PPC, in particular suicide, self-harm, and eating disorder content, on recommended feeds.

20.44 There are technical challenges to moderating content harmful to children at scale, particularly due to the lack of granularity of content classification technologies and where user applied tags purposefully disguise harmful content. Where PPC is not quickly identified after upload, there is a risk that it will be recommended to children and, potentially, in significant quantities.[732] By focusing on content 'likely to be PPC', and content which may have been flagged as potential PPC but is awaiting further review, and enabling services to use all relevant available information to indicate content that is likely to be harmful to children, our aim is that this proposed measure mitigates against these risks. The precautionary approach we are proposing that services take would ensure that children are consistently not recommended content likely to be PPC, thereby reducing their exposure to PPC that would otherwise be recommended to them.

20.45 Some stakeholders have recognised the approach of restricting recommendations (as opposed to restricting the content) as a proportionate mechanism for striking a balance between freedom of expression and safety, given the limitations presently of accurately identifying some types of harmful content at scale. YouTube states that their policy to reduce recommendations of what they describe as borderline content and misinformation, while still allowing users to access all videos that comply with their Community Guidelines: 'strikes a balance between maintaining a platform for free speech and living up to our responsibility to users'.[733] In their response to Ofcom's 2022 Call for Evidence on Illegal Harms, the Alan Turing Institute told us that 'limiting the spread and visibility of content is an option which may reduce harms whilst only having a moderate Impact on free speech'.

20.46 Many U2U services already take steps to prevent some categories of content from being recommended to users and, in some cases, specifically to children.[734] For some services, this

---

[729] Beat response to our 2023 Call for Evidence: Second phase of online safety regulation.
[730] The Coroner's Service, 2022. Regulation 28 Report to Prevent Future Deaths. [accessed 28th October 2022].
[731] For more detail please refer to the Governance, systems and processes sub-section (7.11) of Volume 3 in this consultation
[732] For more detail please refer to the Governance, systems and processes sub-section (7.11) of Volume 3 in this consultation.
[733] The YouTube Team, 2019. Continuing our work to improve recommendations on YouTube. YouTube Official Blog. 25 January. [accessed 17 April 2024].
[734] Meta, 2024. New Protections to Give Teens More Age-Appropriate Experiences on Our Apps Meta Newsroom, 9 January. [accessed 18th April 2024].; TikTok (Keenan C.), 2022. More ways for our community to enjoy what they love [accessed 18th April 2024].

involves attaching labels, tags, instructions, or additional information to content to enable the recommender systems to manage the prominence of that content accordingly.

20.47 We understand that some U2U services with a recommender system have policies aimed at restricting recommendations for what they describe as 'borderline' content, which is content that does not violate the service's content policies (and therefore is not subject to removal) but comes close to being violative, as well as other types of problematic content.[735] For example:

a) Meta claims to take measures to avoid recommending certain types of harmful content to children across Instagram and Facebook. The aim of these measures is to make it more difficult for children to come across potentially '*sensitive*' content, which includes suicide, self-harm and eating disorder content, even if it's shared by someone they follow.[736] Children are defaulted into the most restrictive settings of Meta's content recommendation controls. Existing approaches differ across Meta owned products. Facebook says it takes action to reduce the distribution of content that may either be 'problematic or low quality' and that it 'may reduce the distribution of "borderline" content'.[737] Instagram says that it uses technology to detect both content and accounts that don't meet their Recommendation Guidelines, to help it avoid recommending harmful content.[738]
b) Pinterest maintains a list of sensitive terms which is used to prevent content from appearing in recommendations where it may violate its policies, including terms associated with self-harm, suicide, eating disorders and drug abuse, which indicate content is likely to be PPC.[739] [740]
c) TikTok maintains content eligibility standards for the For You Feed and restricts recommendations of categories of content that are not permitted by the Terms of Service.[741] YouTube also recently announced that where an account is registered to a user 13-18-years-old, they will limit the recommendation of 'content that compares physical features and idealises some types over others, idealises specific fitness levels or body weights, or displays social aggression in the form on non-contact fights and intimidation'. [742]
d) In their response to our 2023 CFE, Twitter (now X) says that neither the Following tab or the For You tab permits "sensitive content or inappropriate advertising" to be surfaced for known under 18 accounts.[743]

---

[735] Ofcom, 2023 Content moderation in user-to-user online services.
[736] Meta, 2024. New Protections to Give Teens More Age-Appropriate Experiences on Our Apps Meta Newsroom, 9 January. [accessed 18th April 2024]
[737] Meta, 2023. Types of content that we demote [accessed 18 April 2024]
[738]Instagram, (no date). Recommendations on Instagram. [accessed 17 April 2024]
[739] Pinterest response to our 2023 CFE Call for Evidence: Second phase of online safety regulation.
[740] Where the provider prohibits PPC in its terms and conditions for the service, it should consider whether content that is likely to be PPC is in breach of those terms, and, if it is, swiftly take the content down. Where the provider does not prohibit PPC in its terms, and it has identified content that is likely to be PPC, it should further moderate the content. If the content is PPC, the provider should swiftly action that content so as to prevent children from encountering it e.g. by age-gating it.
[741] TikTok, 2023. For You feed Eligibility Standards. [accessed 17 April 2024].
[742] Beser, J. 2023. Continued support for teen wellbeing and mental health on YouTube - YouTube Blog. YouTube Official Blog. November 2nd 2023. [accessed 17 April 2024]. Note: The changes described in the Blog have not yet been rolled out in the UK.
[743] Twitter (now X) response to our 2023 Call for Evidence: Second phase of online safety regulation.

e) Instagram says that it "avoid(s) making recommendations that may be inappropriate for younger viewers".[744]

f) According to Tiktok, when they detect that a video contains mature or complex themes, a maturity score will be allocated to the video to help prevent those under 18 from viewing it across the TikTok experience.[745] Tumblr labels sensitive content and has a user-facing feature which enables users to label sensitive content e.g. that which contains nudity or substance abuse. Any content that has a label on it is not surfaced to under 18s.[746]

20.48 The examples of current practice have been included as an indication that this measure is technically feasible but are not an endorsement of current practice, nor an assessment of implementation.

20.49 The proposed measure RS1 goes further than these examples by recommending that services use all relevant available information to inform the recommender system of content that is likely to be PPC which should then be removed for children.

Rights assessment

20.50 This proposed measure recommends services design their recommender systems to filter out content likely to be PPC from children's recommended feeds. We expect that this measure should help ensure that children are prevented from encountering PPC in their recommended feeds. As set out above, evidence shows that recommender systems heighten the risk of children being exposed to harmful content and that they are a key pathway for children to encounter PPC, particularly to encountering such content repeatedly and/or in large volumes, which risks giving rise to cumulative harm. The consequences of such exposure can include significant harm to children's physical, mental, or emotional wellbeing.[747]

20.51 This measure may have a potential impact on the rights of users (including both children and adults)[748] to privacy (Article 8 of the ECHR), freedom of religion and belief (Article 9 of the ECHR), freedom of expression (Article 10 of the ECHR) and freedom of association (Article 11 of the ECHR). It may also have a potential impact on service providers' rights to freedom of expression. We have therefore considered the extent to which the degree of interference with these rights is proportionate.

20.52 In considering the degree of the potential impact on users' and services providers' rights and whether it is proportionate, we have taken as our starting point the requirements of the Act. The children's safety duties set out in the Act require providers of U2U services to use

---

[744] Instagram, (no date). Recommendations on Instagram. [accessed 17 April 2024].
[745] Keenan, C. TikTok, 2022. More ways for our community to enjoy what they love. 13 July. [accessed 17 April 2024].
[746] Tumblr, (no date). Community Labels. [accessed 17th April 2024].
[747] For more information, see the draft Children's Register of Risks (section 7). In particular, see sections relating to pornographic content, content promoting eating disorders, content promoting suicide, content promoting self- injury.
[748] Adult users also include those who are operating on behalf of a business, or accounts that might also be concerned with other entities, such as charities, as well as those with their own, individual account. Both corporate and individual users can benefit from the right to freedom of expression, and we acknowledge the potential risk of interference with the rights of these users to freedom of expression, in addition to the rights of children and adults as individuals.

proportionate systems and processes to prevent children from encountering PPC.[749] By limiting children's exposure to content likely to be PPC in this way, the proposed measure will seek to secure adequate protections for children from harm, in line with the legitimate aims of the Act. It also aims to secure that a higher level of protection is provided to children who are using the service than adults. Preventing children from encountering PPC acts to prevent the harmful consequences of such content that can be inflicted on them. We therefore consider that a significant public interest exists in measures which aim to prevent children from encountering PPC. This substantial public interest relates to the protection of children's health and morals, public safety and, in particular, the protection of the rights of others, namely child users of regulated services.

20.53 As explained in Section 15, Age Assurance we are recommending that services in scope of this proposed measure also use highly effective age assurance to identify child users who should benefit from the protections offered by this proposed measure (see Age Assurance Measure AA5). We discuss the rights impact we expect to arise in relation to use of age assurance in that section and we do not consider them separately here.

Freedom of expression and association

20.54 As explained in Section 2 (Volume 1) which sets out the legal framework, Article 10 of the ECHR upholds the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. The right to freedom of expression is a qualified right. Ofcom must exercise its duties under the Act in light of users' and services' Article 10 rights and not restrict this right unless it is satisfied that it is necessary and proportionate to do so.

20.55 To the extent that this measure restricts children's ability to access content that is PPC and adults' and other users' ability to share such content with children via recommender systems, we consider that this is justified in line with the duties of the Act, as the benefits of the protections on children should outweigh the restriction on other users' rights to share this type of content with children. In addition, we consider that filtering out PPC from recommended feeds, where identified, is the minimum necessary for services to comply with the duties in the Act – we discuss this in more detail under Measure CM1 in Section 16, Content Moderation.

20.56 However, with this proposed measure, potential interference with users' rights to freedom of expression arises where the service provider restricts children's access to content it considers likely to be PPC, and not only to content that the service provider has determined is PPC in line with their content moderation processes. However as set out above, recommender systems are a key pathway for children encountering PPC and, if content likely to be PPC was not filtered out of children's recommended feeds until confirmed to be PPC, there is a high risk that children will still be encountering PPC. Therefore, we expect that services take a precautionary approach when deciding what content should be filtered out of children's recommended feeds and have provided services some flexibility as to the method for how services can become aware of content likely to be PPC.

20.57 In general, we consider that any interference with users' right to freedom of expression (including of both children and adults) as a result of this measure, and services' rights to impart information to their users in the way that they think most effective, would be limited. This is because the proposed measure does not involve services taking any steps in relation

[749] Section 12(3)(a) of the Act 2023

to all users but only in relation to children and only in relation to children's recommended feeds (see further Age Assurance Measure AA5 in Section 15). This means that this measure does not impose any restrictions on adult users who wish to create and share PPC (for example pornographic content) with users other than children, and such content may still be actively recommended to adult users, provided that children are unable to encounter it in their feeds.

20.58 By focusing our proposed measure on content actively recommended to children, we have sought to address the particular risks posed by exposure to PPC in their recommended feeds in a proportionate way. Therefore, we are significantly limiting the impact on both adults' and children's rights to freedom of expression. While there will be some unavoidable impacts on the types of content children can see on their recommended feeds, which we consider to be needed to protect them from harm as explained above, children may still be able to access such content in other parts of the service (i.e. outside recommended feeds), while it may be awaiting further moderation and depending on other safety measures services have in place.

20.59 We acknowledge that there could be unintended impacts on users' rights to freedom of expression (including those of children and adults) as there is a potential risk that there may be cases where content that is not likely to be PPC, including content that may not be harmful to children is flagged as likely to be PPC and removed from children's recommended feeds as a result of this measure (for example, due to inaccurate labelling).[750] If this happened, for example because service providers decided to take a very broad approach to deciding what to classify as content likely to be PPC, this could mean that children might not become aware of content that would potentially benefit them. Such a consequence could also have a potential unintended impact on other users' ability to share their content with child users. However, given that this would only mean such content would not be actively recommended to children, but they could still in principle access such content on other parts of the service, again we consider any impact on both children's and adults' users freedom of expression rights in this regard to be limited.

20.60 In addition, we consider this risk of misclassification of non-harmful content as content likely to be PPC to be mitigated by the fact that this measure recommends services to use all available information they have to consider and determine whether a particular piece of content is likely to be PPC. Furthermore, we would expect that once content has been flagged as likely to be PPC, in most cases it would then be subject to content moderation in line with the provider's content moderation policies and CM1 in Section 16. If such content is determined not to be PPC following further moderation, we would expect that it would no longer be necessary to remove such content from children's recommended feeds, and it could be reinstated.

20.61 We acknowledge that impacts on freedom of expression as outlined above could, in principle, arise in relation to the most highly protected forms of speech, such as religious or political expression, and in relation to kinds of content that the Act seeks to protect, such as

[750] We recognise that classifiers are not error-proof: they may fail to detect some violative items ('false negatives'), particularly for certain types of violation, such as harassment, where assessing whether content is violative requires an understanding of context and nuance; and they may also wrongly remove items that are not violative ('false positives'). Ofcom, 2023. Content moderation in user-to-user online services.

content of democratic importance and journalistic content. However, we consider there is unlikely to be a systematic effect on these kinds of content: for instance, such content would be unlikely to be particularly vulnerable to being wrongly classified as content likely to be PPC. In addition, we have provided examples of types of content, including protected forms of speech (such as content of journalistic importance), in our Guidance on Content Harmful to Children, Section 8, Volume 3, which we encourage service providers to have regard to in implementing this measure.

20.62 We expect services in scope of this measure would also need to have in place highly effective age assurance to ensure this measure is targeted at child users, for the reasons explained in Section 15, Age Assurance in this Volume. This should mitigate the risk of this measure being applied to adult users in error. While this is not a requirement of the measure, we acknowledge that a greater interference with users' rights (particularly adults' rights) could arise if the service provider chose to apply this measure in a way that meant content likely to be PPC would be filtered out from the recommended feeds of all users, not just child users. In this case, services could also be restricting adult users' access to certain types of content which is not required under the duties in the Act, and might also not be harmful, or might be less severely harmful, to them. However, it remains open to services as a commercial matter (and in the exercise of their own right to freedom of expression) to decide what forms of content to allow or not to allow to be promoted on recommended feeds on their service so long as they comply with the Act. Services have incentives to meet their users' expectations in this regard.

20.63 We recognise that more significant impacts to users' rights to freedom of expression and association could arise if services choose to withdraw their recommender systems, or to withdraw the service from the UK market entirely (for instance, if the recommender system is integral to the service's business model) due to the costs of the proposed recommender systems changes (together with the cost of implementing highly effective age assurance under the related Age Assurance Measure AA5 in Section 15). However, we have given service providers flexibility as to how to implement this measure in a way which minimises the costs so far as possible. In addition, we consider it unlikely that most services in scope of this measure would take these steps. We expect that many services will retain commercial incentives to enable users in the UK (both children and adults) to continue to use the service and would not typically see very large reductions in user engagement due to this measure. Therefore, we would expect that UK users will still have a large range of services from which they can benefit, even if their choice were to be somewhat more limited than it is currently.

20.64 We consider the implementation of this measure could also have positive impacts on freedom of expression and freedom of association rights of children as we expect it will result in limiting children's exposure to PPC content which would result in safer spaces online where children may feel more able to join online communities and receive and impart (non-harmful) ideas and information with other users, providing significant benefits to children.

20.65 For the reasons set out above, we consider that the impact of the proposed measure on users' rights to freedom of expression to be limited, and likely to go no further than needed to secure the positive benefits to children that are intended through this measure. We consider that the impact on users' rights to freedom of expression and association is therefore proportionate.

20.66 The proposed measure may also have an impact on service providers' rights to freedom of expression as, to the extent that they do not already choose to restrict children's exposure to content that is likely to be PPC, services would need to put in place steps to ensure that it is appropriately dealt with in line with this measure. However, most of this impact arises from the duties placed on services under the Act by Parliament which imposes the duty to prevent all children from encountering PPC, and we consider any additional impacts associated with filtering out content likely to be PPC from children's recommended feeds to be limited. Taking this, and the benefits to children into consideration, we consider that the impact on service providers' rights to freedom of expression and association is therefore proportionate.

Privacy

20.67 This proposed measure applies to recommender systems which are made up of many different algorithms. The algorithms used in recommender systems are designed to identify relationships between content and users by analysing content features and characteristics, as well as learning about users' preferences (typically by means of machine learning). Additionally, content moderation is one of the processes that provides recommender systems with relevant available information about the nature of content and other types of metadata. The content lifecycle from upload, content moderation, recommendation, and interaction involves some degree of processing personal data of individuals, including children. It will therefore impact on users' rights to privacy and their rights under data protection law. The degree of interference will depend on the extent to which the nature of their affected content and communications is public or private, or, in other words, gives rise to a legitimate expectation of privacy. However, we have not identified any specific potential impacts connected with restrictions on children's or adults' private communications, as by their nature, recommender systems would generally only promote content that is widely publicly available, rather than private communications. We therefore consider that it is less likely that this measure would lead to review of content or communications in relation to which individuals might expect a reasonable degree of privacy (though we acknowledge this could still occur).

20.68 The degree of impact will also depend on the extent of personal data about individuals that is processed. However, the proposed measure (RS1) does not specify that service providers should obtain or retain any specific types of personal data about individual users as part of any content moderation processes they undertake as part of their content recommender systems, and we consider that service providers can implement the measure in a way which minimises any potential impact on users' right to privacy. In processing any users' personal data for the purposes of this measure, services would need to comply with relevant data protection legislation. This means they should apply appropriate safeguards. Insofar as services use automated processing to implement this measure, we consider that there is a potentially more significant impact on users' rights to privacy, especially if they are unaware that their personal data will be used in this way. Services should refer to ICO guidance[751] to determine whether the processing is solely automated i.e. has no meaningful human involvement, and results in decisions that have a legal or similarly significant effect on users.[752]

[751] ICO (no date) Data protection principles - guidance and resources and Content moderation and data protection [accessed 25 April 2024]

[752] In which case Article 22 UK GDPR requirements are likely to apply.

20.69 For the reasons set out above, we do not consider there to be any material impact on users' rights to privacy and the measure to be proportionate on that basis, particularly considering the benefits to children that it would secure.

Impacts on services

20.70 We consider separately below the direct costs of modifying the service to implement the proposed measure, costs related to age assurance, and the potential for an indirect cost to service providers resulting from lost revenue.

20.71 Table 20.1 below presents quantified estimates of direct costs, based on the assumptions summarised in this sub-section. Although we have drawn on available evidence and expert input, our quantitative estimates of costs should be interpreted as indicative. Real world costs will depend on the specific recommender systems and associated systems used by service providers. Table 20.2 below presents illustrative costs of age checks which is discussed in more detail in Section 15, Age Assurance.

Table 20.1: Summary of direct cost estimates

Activity	One-off implementation cost	Ongoing annual cost
Implementing the measure	£13,000 to £80,000	£3,000 to £20,000
Linking Highly Effective Age Assurance to the measure	£9,000 to £36,000	£2,000 to £9,000

Source: Ofcom analysis

Table 20.2: Illustrative cost estimates of age checks via third-party age assurance providers[753]

Service size	Existing UK user base	New users each year	Age assurance for existing users	Age assurance for new users (annual ongoing cost)
Smaller service	100,000	10,000	£5,000 to £20,000	£1,000 to £2,000
Larger service	7,000,000	70,000	£350,000 to £1,400,000	£4,000 to £14,000

Source: Ofcom analysis. Note that the above estimates are based on age checks being conducted for all users, which is likely to be an upper bound and may overestimate costs, as we explain below under 'Costs related to age assurance'.

Direct costs of implementation

20.72 We understand that there will be costs associated with adjusting systems to filter out content likely to be PPC from the recommender feeds of children. We believe that service providers that use content recommender systems are likely to have access to specialist engineers and computing infrastructure to modify their recommender system. However, implementing this measure will require the diversion of some of these existing resources, or additional resources, and therefore will have a direct cost to the service provider. We set out

[753] For further detail on age assurance cost analysis see Section 15 and Annex 12.

below our understanding of the activities and associated costs on service providers of the steps described in the explanation of the measure that may need to be implemented to follow this measure.

20.73 Service providers will incur costs in relation to: (1) using relevant available information to identify content likely to be PPC, (2) making the signal available to the recommender system, and (3) modifying the recommender system to filter out content likely to be PPC from content recommended to children.

1. **Using relevant available information to identify content likely to be PPC.** As set out in Definition Box 1 above, there are different sources of relevant available information that a service provider could use to indicate that content is likely to be PPC. We are not prescriptive about the types of information used or how this is used, and therefore specific approaches and associated costs may vary. Service providers have flexibility in determining how these sources of relevant available information indicate that content is likely to be PPC. Service providers could set thresholds from different sources of relevant information. One approach might be for providers to combine sources of available relevant information to assess if content is likely to be PPC, for instance using a machine learning model to incorporate different sources. We understand that this could be similar to existing models that service providers may use to determine if, and with what priority, content should be queued for human moderation.
2. **Making the signal available to the recommender system.** To implement this measure, the recommender system of a service provider needs to receive information that content is likely to be PPC in a way that can be actioned. We understand that one way of achieving this is by attaching safety labels to content that is likely to be PPC based on relevant available information (for example through content moderation and user reports). Safety labels serve as signals available to the recommender system. We understand that as recommender systems need to receive information about content to inform recommendation decisions, regardless of our proposed measure, that this is feasible.
3. **Modify the recommender system to filter out content likely to be PPC from content recommended to children.** Service providers would then need to ensure the recommender system can action signals about content that is likely to be PPC. We understand that ways that this could be achieved include adding additional filters so that the content is removed from the content pool for children, or re-programming the component of the recommender system that is relevant for re-ranking to filter out content likely to be PPC from children's recommended content. This might be implemented by redesigning the re-ranking algorithm so that it can assign a default relevancy score of 0 to content likely to be PPC.

20.74 We understand that the main cost in implementing this measure would be labour input from software engineers, machine learning teams, and data specialists. Service providers will likely need to test their recommender system once it has been redesigned, and additional time may be required to conduct these to ensure effectiveness.

20.75 We estimate that for a service provider to undertake the three activities above could require a one-off direct build effort of approximately 6-18 weeks of labour time split across roles including software engineers, machine learning engineers, and data scientists. We have assumed that this time is matched with an equal amount of non-software engineering time (e.g. project management, legal, trust and safety). Using our assumptions on labour costs

required for software engineering work set out in Annex 12, we estimate that one-off direct costs could be in the region of £13,000 to £80,000.

20.76 The cost of this measure for a given service provider will be impacted by the existing design of the recommender system. We consider that costs will be higher for services with more recommender systems operating, and where systems are complex (for instance serving more users in more languages) but do not already have a mechanism for limiting the prominence of certain types of content. As set out in the 'Effectiveness' sub-section above, many service providers have already designed their recommender system to ensure that certain types of content are not recommended. It may be more straightforward for service providers to implement this measure if they already have the infrastructure in place to remove content from recommendation feeds and only require more minor adjustments.

20.77 While not specifically recommended by our measure, we also consider that where a service provider chooses to build a machine learning model (to combine relevant available information to assess if content is likely to be PPC and use this as a signal for the recommender system) there are potential costs associated with model training, in addition to the design costs, such as compute costs. We expect these costs would vary with the size of a service and could be in the tens of thousands for some businesses, but we do not have sufficient information to make a quantitative estimate of them.

20.78 We consider that there may also be additional business oversight and coordination costs associated with changing products. Larger businesses may use more complex processes for system changes and face significant review, communication and legal processes to implement changes to their services. Such businesses may incur higher costs than the indicative figures presented in this section. We would expect the oversight and coordination costs to be largely correlated with the size of the company, but do not have sufficient information to be able to quantify these.

20.79 In addition to the implementation costs, we would expect a service provider to incur ongoing costs including the maintenance costs of this measure to ensure that it continues to function as intended. There may also be ongoing costs of an extended product management cycle where providers may have additional objects to consider as part of ongoing management, for instance where there are additional variables to observe in terms of how the measure is performing. In line with our standard cost assumptions set out in Annex 12, this could be approximately 25% of the initial set-up costs, which equates to £3,000 to £20,000 per year.

20.80 Measure RS2 in this section is likely to involve similar activities and teams to carry out the necessary work to implement this Measure RS1. For instance, identifying and potentially combining relevant available information that can be used to assess if content is likely to be PPC or PC may be a similar activity. Ensuring that the recommender can follow instructions according to safety signals may be a common task to both measures. Finally, we understand that both measures could be implemented through changes to the re-ranking component of the recommender system. We believe that the implementation of Measures RS1 and RS2 would likely be undertaken jointly by service providers in scope of both measures, and a result there could be some synergies when services are making these changes simultaneously. While these synergies could be significant in the case of some service providers, there is a high degree of uncertainty about the degree and variation of the overlap of costs of the two measures for different services. Therefore, we have not

quantified any estimated cost reduction for service providers implementing these measures simultaneously.

Costs related to age assurance

20.81 Service providers in scope of this measure should apply highly effective age assurance to target this measure at children, although this is not specifically recommended. The costs of highly effective age assurance are covered in Section 15 and our discussion of Measures AA5 and AA6.

20.82 As we explain in that chapter, where service providers opt to use third-party providers of age assurance solutions, there will be some one-off costs to understand our highly effective age assurance guidance, assess and choose appropriate third-party age assurance methods, and integrate these. In many cases, the bulk of costs may come from third-party provider fees, which are typically linked to the volume of age checks conducted and could amount to between 5p and 20p per age check. As set out in Table 20.2 above, for a service with 100,000 users, age assurance for existing users could cost between £5,000 to £20,000, and if the service had 10,000 new users each year this could result in an ongoing annual cost of £1,000 to £2,000 for these users. For larger service providers, costs may be significantly higher. For a service provider with 7,000,000 users, age assurance for existing users could cost between £350,000 to £1,400,000, and if the service had 70,000 new users each year this could result in an ongoing annual cost of £4,000 - £14,000 or these users.

20.83 Our Age Assurance Measures AA5 and AA6 provide flexibility to service providers and do not necessarily require age checks on all users. For example, a service may choose to conduct age checks only where users request to access an unfiltered recommender feed. Therefore, the cost to the provider depends on the approach taken – including where age assurance is applied in the user journey – and may also depend on how motivated its adult users are to remove recommender system filtering. The values in Table 20.2 are therefore more likely to represent an upper bound and the age assurance costs could be significantly lower if age checks are only performed on a subset of users.

20.84 There will also be some limited costs to update relevant parts of the user interface or settings related to the interaction between age assurance and the recommender system (for example, messages to inform users that they need to complete an age check to unlock an unfiltered recommender feed).

20.85 We understand that there may also be costs to integrate the age information from the highly effective age assurance into the recommender system. This might take the form of adding an age-based layer, and we consider this could be around 4 – 8 weeks of engineering time, with an equal amount of non-software engineering time (e.g. project management). Using our assumptions on labour costs required for this type of work set out in Annex 12, we would expect the one-off direct costs to be up to the region of £9,000 to £36,000, and annual maintenance of £2,000 to £9,000.

20.86 Costs related to age assurance would also apply to Measures RS2 and RS3 in this section, but they would only need to be incurred once if a service provider is in scope of more than one of the proposed Recommender System measures.

Indirect costs to services

20.87 This measure may have an indirect cost on service providers to the extent it might impact their business model (revenue model and growth strategy). Certain business models,

including advertising and subscription models, generate revenue for a service in proportion to the number of users and/or their engagement.[754]

20.88 We expect this measure to result in children not being recommended PPC content that would have been recommended to them otherwise. To the extent this leads to a loss of revenue for a provider, we consider this entirely justifiable given the importance of preventing children from encountering PPC.

20.89 However, content likely to be PPC may include some content that is not in fact PPC and not harmful to children, which would not be recommended to children because of this measure. This is a potential unintended consequence of this measure and the recommended precautionary approach of not recommending content that has not completed a final determination. We have considered whether this may have an impact on users' engagement with the service provider where non-harmful content is not recommended but would have been engaging for the user. However, where content is not recommended as it is likely to be PPC, we expect that alternative content will be recommended to those users. We consider that there is a wide variety of content other than content that is likely to be PPC that engages and benefits children, so it is not clear that removing content likely to be PPC would necessarily materially reduce children's engagement with a service provider.

20.90 We also believe that there is a potential countervailing indirect benefit to service providers which reduce children's exposure to harmful content through recommender systems. Firstly, some users may reduce usage of services where they encounter harmful content. Secondly, service providers may face commercial pressure from businesses or advertisers who do not want to be associated or positioned close to harmful content within recommender feeds. The Business models and commercial profiles, Section 7.12 in Volume 3 sets out how these reputational risks can negatively affect a service provider's revenue in the long run.[755] These effects may to some extent reduce the negative impact to providers of this measure.

Which providers we propose should implement this measure

20.91 We propose to recommend this measure for providers of all U2U services likely to be accessed by children that have a content recommender system (as set out in 'An explainer: What is a recommendation system?' above), and which are medium or high risk for at least one kind of PPC. As explained in Section 15, the Age Assurance section of this Volume, we consider that services providers in scope of this measure should apply highly effective age assurance to target the measure at children and achieve the intended effect – see Measure AA5. In assessing the proportionality of this measure for different kinds of service providers, we therefore consider the impacts of both this measure and the related age assurance measure in the round.

20.92 Recommender systems are a key pathway for children to encounter all forms of PPC. They can introduce children to this content for the first time and can facilitate repeated engagement with harmful content, leading to children experiencing cumulative harm. This can occur even on service providers where PPC is prohibited, given the challenges of

[754] In September 2021, The Wall Street Journal described how one company targeted an app at teens despite having conducted research that showed evidence of harm, specifically in relation to body image, associated with the service. The Wall Street Journal, 2021. Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show. [accessed 25 April 2024].

[755] For more detail, please see Business models and commercial profiles set out in volume 3, sub-section 7.12.

moderating content at scale, and results in this content often appearing in recommender feeds. The proposed measure would mean that providers take a precautionary approach in not recommending content to children when there are sufficient indications that it may be harmful to them, even when this is not yet confirmed. We therefore expect the measure to make a significant contribution to protecting children from harm.

20.93 We consider that this measure should apply to service providers whose risk assessment indicates that children face a medium or high risk for at least one kind of PPC because this is where this measure will create material benefits by helping prevent children encountering this content.

20.94 The estimated costs of this measure for service providers can be significant. We recognise the possibility that a minority of small businesses in scope of this measure could struggle to carry this cost. Providers may be discouraged from offering recommender systems, and where these are integral to business models, it could discourage some service providers from serving UK users. This could harm users who benefit from accessing these functionalities and even services.

20.95 However, in most cases we consider that the costs to service providers will vary depending on the complexity of each provider's recommender system. Costs will be higher for service providers with more complex recommender systems, including where these cater for large volumes of users and multiple languages, and we believe these providers will typically have greater capacity to implement changes. We have designed this measure to allow some flexibility in how it is implemented, enabling service providers to manage their costs accordingly. We also believe that providers running recommender systems have the necessary technical capabilities to implement this measure. Our assessment of costs and judgement of proportionality is based upon considerations that service providers with these systems already have a level of technical maturity to allow their recommender systems to receive and action relevant information.

20.96 The related age assurance costs are expected to largely scale with size of service provider and may also scale with the level of risk. The riskiest service providers are more likely to be filtering large volumes of content likely to be PPC from their recommender feeds, which may motivate a greater proportion of adult users to conduct an age check and unlock this content in their feeds. While costs may be higher for such providers, the benefit to children's safety from this measure is also higher on the riskiest service providers.

20.97 Overall, we recognise the measure imposes material costs that could lead to some loss of user choice if small services struggle to shoulder the burden of this measure. We nonetheless consider it proportionate to apply it to all services who have a recommender system and are at medium or high risk for at least one kind of PPC (regardless of size) given our view of the effectiveness of the measure and of the important role played by these systems in exposing children to harm related to PPC.

20.98 We do not recommend this measure for providers where the risk of PPC is low, because the measure would have limited benefits for children's safety, if any, while its impacts on service providers and adult users would be material.

20.99 We therefore propose to recommend this measure to all U2U service providers likely to be accessed by children (regardless of size) that have a content recommender system, and which are medium or high risk for at least one kind of PPC.

Provisional conclusion

20.100 Given the harms this measure seeks to mitigate in respect of all kinds of PPC, as well as the risks of cumulative harm that recommender systems pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU F1 in Annex A7.

Measure RS2: Recommender systems to reduce the prominence of content likely to be PC

Explanation of the measure

20.101 Service providers should design their recommender systems with the relevant capabilities to significantly reduce the prominence of content that is likely to be PC on children's recommender feeds. Whereas under RS1 we are proposing that service providers consistently filter out content likely to be PPC from the recommender system, this measure involves service providers consistently reducing the prominence and visibility of content likely to be PC,[756] for example by downranking the content in the recommender system.

20.102 We propose that service providers should design their recommender systems with the relevant capabilities to significantly reduce the prominence[757] of content that is likely to be PC (this includes content that has been confirmed as PC through content moderation) on children's recommender feeds. Reducing prominence would minimise the visibility of content likely to be PC on users' recommended feeds. This would mean content likely to be PC is difficult to organically encounter by children.

20.103 We propose to extend this measure to additional categories of non-designated content, such as body image content and depressive content, where the volume and frequency of exposure plays a part in amplifying the harm. However, this would be subject to the outcome of the consultation on our proposals to classify body image and depressive content as non-designated content.[758]

20.104 **Downranking** is an important function carried out during the re-ranking phase of the recommender system. At the re-ranking stage a recommender system typically adjusts the final list of content recommendations based on a variety of factors. These factors can include efforts to promote content diversity, variety, and age-appropriate content at scale. Downranking refers to the process of reducing the visibility of content that does not violate a service provider's terms of service but may be considered inappropriate for widespread dissemination (e.g., content that is potentially misleading or low-quality content).

---

[756] We expect services to take appropriate action against bullying content in relation to Section 16, Content Moderation and this measure, however given limited evidence of the connection between bullying content and recommender systems a service would not be in scope of the measure if the children's risk assessment does not identify any risk of PC other than bullying on the service. If we receive more evidence of the connection between bullying and recommender systems, we may be able to revise this position.

[757] By prominence, we refer to the placement of the content in recommender feeds, particularly how visible and highlighted content is for a user (or how easily it can be organically encountered). For example, prominent content may be encountered as one of the top 10 or N items, and downranked content might be encountered after the 100th or Nth item.

[758] Please see the draft Children's Register of Risks and Harms Guidance in Volume 3 for further detail on depressive and body image content proposals.

Downranking content can also be an outcome of user signalling negative sentiment towards a particular category of content (see Measure RS3). Alongside downranking, certain items that service providers consider desirable may be deliberately given more prominence (e.g. popular events or trending topics). We are recommending that services reduce the prominence of content that is likely to be PC for users that are children.[759]

Definition Box 3: What is content likely to be Priority Content (PC)?

Priority Content (PC) is content whereby services have a duty to use proportionate systems and processes to protect children in age groups judged to be at risk of harm from encountering in accordance with the Act.[760] This content includes online abuse and hate, content depicting violence, dangerous stunts and challenges, content encouraging a person to ingest harmful substances and bullying content.[761] Further information about PC can be found in Guidance on content harmful to children in Volume 3, Sections 7.4-7.8.

For the purposes of this measure, content likely to be PC is expected to include:

- Content that has completed moderation and is therefore known to be PC. In practice, this should mean that the content is addressed at the moderation stage as set out in Section 16, the Content moderation for U2U services within this Volume. However, were this to make it through to the recommender pool for children, this should be reduced in prominence for children.
- Content which is undergoing content moderation or has been flagged for moderation.
- Content which has not undergone any form of content moderation but where relevant information indicates that there is a material likelihood of that content being PC.[762]

To achieve a reduction in the visibility of PC U2U services should leverage their existing capabilities to utilise relevant available information as instructions for the recommender system to limit the prominence of content likely to be PC. User reports or other tags applied by users at upload can also provide a signal to a service that the content is likely to be PC. For the purposes of implementing this measure, content likely to be PC includes all categories of PC (including bullying content). (For the purposes of determining whether a service provider is in scope of the measure, risks relating to bullying is excluded.)

[759] As is the case with Measure RS1, the benefits of this proposed measure, in terms of improved protection of children, are partly contingent on the separate Measures AA5 and AA6 as set out in the Age Assurance section 15 within this Volume. This ensures that children are protected by this measure as opposed to applying to all users. Service providers in scope of these measures will incur costs associated with both this proposed measure and the relevant proposed measure described in the Age Assurance. While we discuss these measures separately in the respective sections, we have had regard to the combined costs and benefits in the round as part of our assessment and explanation of how the measure works.

[760] Section 12(3)(b) of the Act.

[761] Section 62 of the Act.

[762] Separately, CM4 in this Volume recommends that services have regard to whether the likelihood that content is PC in deciding which moderation cases should be prioritised. RS2 goes further by requiring action to be taken on "likely to be" content in recommender systems, even though a piece of content does not yet have a final moderation determination.

20.105 To ensure that content that is likely to be PC is captured by the measure and is less prominent in the feeds of children and therefore is less likely to be encountered by children, we recommend that service providers (in scope of this measure) deploy the following changes to their recommender systems:

- For this measure, services should use the relevant available information (see Definition Box 1) resulting from their existing systems and processes to consistently reduce the prominence of content that is likely to be PC in the recommender system so that this content is limited on the recommender feeds of children. As described in Definition Box 1, we expect service providers to already have relevant available information and this can include but is not limited to content metadata, such as tags, labels, or other information that suggests that the content is likely to be PC such as user reports.
- This relevant available information should serve as a signal for the recommender system to reduce the prominence of that content on children's recommender feeds, to protect children in age groups judged to be at risk of harm from content likely to be PC. As set out in the above sub-section 'What are recommender systems and how do they work', we understand that recommender systems are capable of processing a variety of signals about content. Depending on how these signals are processed and actioned by the recommender system, they can have a significant impact on the visibility of that type of content on recommender feeds. These signals can come from content moderation processes in the form of labels and age ratings (e.g. mature, sensitive, violent, misinformation).[763] They can also take the form of keyword detection, where certain user tags might be considered associated with undesirable content. User complaints can also be a signal that results in content likely to be PC being reduced in prominence for children.[764] All of these signals, irrespective of their origin, could act as relevant available information for the recommender system on whether content is likely to be PC and whether the prominence should, therefore, be limited.
- As set out above, one way of doing this may be to downrank content during the re-ranking stage of the recommendation process.[765] Regardless of how services chose to reduce the prominence of PC, this should be designed in a way that overrides any existing engagement patterns (likes, watch time, reshares etc.) on content that is likely to be PC. Positive user feedback on content likely to be PC, including inferred feedback by the child user, should not increase the probability of that content being recommended to children.

20.106 As set out above (see sub-section 'What are recommender systems and how do they work?'), the typical recommender system uses scoring algorithms to curate content that is considered relevant for the user, then re-ranking algorithms to carry out a variety of safety functions to ensure that the content recommended to the user is appropriate. Re-ranking algorithms also carry out important tasks to avoid homogenised content feeds by ensuring diversity of content.[766]

20.107 As we set out for Measure RS1, to allow services flexibility in the kinds of information that the recommender system can use to inform whether content is reduced in prominence, we

[763] Thorburn, L, Bengani, P, Stray, J. 2022. How platform recommenders work. [accessed 12 April 2024].
[764] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.
[765] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.
[766] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.

are not specifying any particular techniques or methods for identifying content likely to be PC. This is consistent with the approach set out in the Content Moderation, Section 16 within this Volume which provisionally recommends that all service providers should operate a content moderation system to swiftly identify and action content that is harmful to children but leaves service providers with flexibility as to how to design and deploy these systems. As such, when service providers deploy this measure, they can use a variety of signals they consider sufficiently indicative of content that is likely to be PC to achieve the outcome of limiting the visibility of PC on the recommender feeds of children.

20.108 We consider that the steps service providers may need to take to do this are: (1) using relevant available information to identify content likely to be PC, (2) making the signal available to the recommender system, and (3) modifying the recommender system to reduce the prominence of this content on children's recommender feeds.

Effectiveness at addressing risks to children

20.109 The purpose of downranking content that is likely to be PC is to make it consistently less visible to children by reducing the frequency of this type of content in relation to other types of items of content. Downranking content can be an effective strategy for significantly limiting the prevalence of harmful content on recommender feeds. This minimises the risk of children unexpectedly coming across this content and incur cumulative harm from continued exposure to this content. Some service providers already leverage downranking as a means of minimising the prevalence and visibility of content that they consider undesirable. Examples include X (formerly Twitter) downranking misinformation[767] and Meta downranking a variety of content types.[768]

20.110 In response to concerns about users being recommended extreme content on YouTube, YouTube announced changes in 2019 to "reduce the spread of content that comes close to—but does not quite cross the line of—violating our Community Guidelines". It claimed that these interventions resulted in a 50% reduction in watch time from recommendations for what they describe as "borderline content and harmful misinformation" and a 70% decline in watch time from nonsubscriber recommendations.[769]

20.111 In accordance with the Act,[770] this measure is designed to protect children from PC by minimising the risk of children unexpectedly encountering and viewing this content. By limiting this risk, this measure can result in downstream benefits by reducing the risk of unintentional engagement with PC. As we set out in the draft Children's Register of Risks, our research reported that recommender systems were generally stated as the main way in which children encountered violent content without seeking it out, largely from strangers on their personalised news feeds. Children said they felt they had no control over the content they were recommended, and therefore seeing more violent content felt inevitable.[771]

---

[767] Roth, Y. and Pickles, N., 2020. Updating our approach to misleading information. X Blog, 11 May. [accessed 18th April 2024.] Note: X is no longer enforcing the COVID-19 misleading information policy as of November 2023.

[768] Meta, 2023. Types of content that we demote. Meta Transparency Center, 16 October. [accessed 18 April 2024].

[769] Goodrow C, 2021. On YouTube's recommendation system. YouTube Official Blog, 15 September. [accessed 18 April 2024].

[770] Section 12(3)(b) of the Act.

[771] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.

20.112 Another study researched how a recommender system actively amplified and directed abusive and hateful content to young people; through increased usage, users were gradually exposed to more misogynistic ideologies which were presented and gamified.[772] As this is the case for violent content, and our evidence base for other forms of PC is emerging, it is reasonable to assume that other forms of PC may also be encountered via recommender systems.

20.113 In summary, by using relevant available information to inform the recommender to downrank content that is likely to be PC, this measure can significantly limit the risk of children encountering PC via recommender systems.

20.114 Our research, outlined below, indicates that it is some industry practice to limit the prominence and visibility of specific kinds of content, however these are not always effectively implemented nor are they targeted at protecting children. Below are some examples of industry efforts to minimise the prominence of content they consider undesirable for widespread dissemination.

a) Alongside prohibiting certain types of content in their Terms of Service, Meta has content distribution guidelines that describe types of content that they consider should be demoted.[773] Content that Meta considers low quality[774] is deliberately limited in distribution across Facebook. Content that is downranked on Facebook includes sensationalist health content,[775] engagement bait,[776] and pages considered to be spam. Meta also redesigned its recommender system to change the way political content is ranked on user's personal newsfeeds. This included moving away from ranking based on engagement and giving more importance to content that is more informative and meaningful to users. In practice, we consider this to involve manually reducing the relevance (i.e., downranking) of popular political content in user feeds.[777]
b) TikTok has several product policies in place to ensure that its recommender system provides users with a variety of content to minimise the risk of rabbit holes and filter bubbles. TikTok does this by interspersing the recommendations users receive with content that falls outside of users' explicit preferences. For example, TikTok's Friends Tab and Following Feeds will generally not recommend two videos in a row made by the same creator. Similarly, TikTok's For You and Live Feeds typically avoid recommending content that has been viewed before.[778]
c) YouTube has a suite of product policies that are designed to promote authoritative content and reduce the prevalence of borderline content and harmful

[772] University College London and University of Kent, 2024. Safer Scrolling: How algorithms popularise and gamify online hate and misogyny for young people. [accessed 28 March 2024].
[773] Stepanov A, 2021. Content Distribution Guidelines. Meta Newsroom. 23 September. [accessed 18th April 2024].
[774] Based on user feedback on the type of content they do not like to see, Meta downrank content they consider to be problematic or low quality. This includes ad farm content, clickbait, spam, sensationalised health posts, and comments likely to be reported. See Meta, 2023. Types of content that we demote. Meta Transparency Center, 16 October. [accessed 18th April 2024].
[775] Meta, 2024 Sensationalist health content and commercial health posts. Meta Transparency Center. (no date). [accessed 18th April 2024].
[776] Meta, 2024, (no date). Engagement bait. Meta Transparency Center. (no date). [accessed 18th April 2024].
[777] Stepanov A and Gupta A, 2021. Reducing Political Content in News Feed. Meta Newsroom, 10 February. [accessed 18 April 2024].
[778] TikTok, (no date). How TikTok recommends content TikTok Official Blog. [accessed 10th April 2024].

misinformation.[779] Between 2015 and 2019, YouTube introduced a series of changes to its recommender system to ensure that non-violative but objectionable content (such as borderline content and harmful misinformation) was reduced in prevalence. In addition to downranking low-quality content, the YouTube recommender system is designed to promote and surface content considered authoritative, high-quality, and factual. These changes to the YouTube recommendation system were accompanied by wider use experience changes to improve the visibility and accessibility of content promoted, such as information panels and news shelves.[780] YouTube also recently announced that where an account is registered to a user 13-18 years old, they will limit the recommendation of 'content that compares physical features and idealizes some types over others, idealizes specific fitness levels or body weights, or displays social aggression in the form on non-contact fights and intimidation'. [781]

20.115 Additionally, some service providers have specific rules around recommendations for children, however these are implemented with varying effectiveness and do not currently focus on the specific function of consistently reducing the prominence of content that is likely to be PC in recommended content for children. For example:

a) TikTok has recently introduced a system to organise content based on thematic maturity.[782] When TikTok detects that a video contains mature or complex themes, for example fictional scenes that may be too frightening or intense for younger audiences, a maturity score will be allocated to the video to help prevent those under 18 from viewing it across the TikTok experience.[783]
b) In response to our 2023 CFE, Meta told us that it adds a warning label to especially graphic or violent content so that it is not available to users under 18.[784] Tumblr has a user-facing feature called 'community labels' which enables users to label content to indicate that it contains e.g. nudity or substance abuse. Tumblr also has the ability to label content itself. Any content that has a label on it is not surfaced to under 18s or to adults who have chosen in their settings not to view such content.[785]

20.116 We believe the above examples of current practice suggest that managing the visibility of specific content that is considered harmful or which may not appeal to users or children specifically is already used by some U2U service providers. This supports our provisional view that the proposals set out in this measure are technically feasible and actionable by providers operating a recommender system. We propose service providers take a precautionary approach by reducing the prominence of content that is likely to be PC, this includes content that has not been confirmed as being PC. If a service provider prohibits all forms of PC then we would still expect this measure to be effective as it would also capture

---

[779] Mohan N, 2022. Inside Responsibility: What's next on our misinfo efforts . YouTube Official Blog, 17 February. [accessed 18th April 2024].
[780] YouTube, 2019. The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation. 3 December. [accessed 18th April 2024].
[781] Beser, J. 2023. Continued support for teen wellbeing and mental health on YouTube - YouTube Blog. YouTube Official Blog, 2 November. [accessed 17 April 2024]. Note: The changes described in the Blog have not yet been rolled out in the UK.
[782] Keenan C, 2022. More ways for our community to enjoy what they love. TikTok Newsroom, 13 July. [accessed 18th April 2024].
[783] Keenan C, 2022. More ways for our community to enjoy what they love. TikTok Newsroom. 13 July. [accessed 18th April 2024].
[784] Meta response to 2023 Call for Evidence: Second phase of online safety regulation.
[785] Tumblr, (no date). Community Labels. [accessed 17th April 2024].

content that is likely to be PC. We anticipate this will mean that the prominence of content likely to be PC appearing in recommendations to children will reduce from current practice across in scope service providers.

20.117 The examples of current practice cited above have been included to indicate that these measures are technically feasible but should not be seen as an endorsement of these measures being implemented effectively nor an indication of compliance with the proposed measure.

Rights assessment

20.118 This proposed measure recommends that service providers reduce the prominence of content likely to be PC in children's recommended feeds. We expect this to result in a significant reduction of visibility of this content on children recommended feeds. As set out above, evidence shows that recommender systems heighten the risk of children being exposed to harmful content.

20.119 In implementing this measure, there is a potential impact on the right of users and service providers, in particular, their rights to privacy (Article 8 of the ECHR) and their rights to freedom of expression (Article 10 of the ECHR), as well as a potential impact on service providers' rights to freedom of expression. We have considered the extent to which the degree of interference with these rights is proportionate. In doing so, our starting point is to recognise that the children's safety duties set out in the Act require providers of U2U services to use proportionate systems and processes designed to protect children in relevant age groups from encountering PC.[786] The Act also aims to secure that a higher level of protection is accorded to children who are using a service than adults. By reducing children's exposure to content likely to be PC in this way, the proposed measure will seek to protect children from the harmful consequences of such content that can be inflicted on them, particularly from encountering such content repeatedly and/or in large volumes, which risks giving rise to cumulative harm. These consequences can include harm to children's physical, mental or emotional wellbeing. We, therefore, take the view that a substantial public interest exists in measures which aim to protect children from encountering PC.

20.120 We consider that the impact on users' and service providers' rights to freedom of expression will be very similar to those set out in relation to Measure RS1 above, and we therefore adopt that reasoning in respect of this Measure (RS2) so far as applicable. We note however that all things being equal, the impact on users' rights to freedom of expression would tend to be more limited than for Measure RS1 on the basis that we are only proposing that content likely to be PC is limited in visibility and prominence in children's recommended feeds, which is a less restrictive action than filtering out as proposed for PPC. We consider this less restrictive action is justified as the Act recognises that children require stricter protections against the most harmful content on U2U services (e.g. it requires steps to be taken to prevent children of all ages from encountering PPC), compared to the requirements to protect children in relevant age groups from encountering PC.

20.121 On the other hand, we also recognise that the duties in the Act recognise that some age groups may require less, or no, protection from some forms of PC, as it may be significantly less harmful to them. However we are not proposing at this time measures which are tailored at a particular age, for the reasons discussed under 'Age groups' in the Age

[786] Section 12(3)(b) of the Act.

Assurance section, in particular due to limited evidence on the technical capability for services to place children into age groups below the age of 18 and due to the limited evidence in linking specific PC harms to different age groups.

20.122 We also note that the severity of impacts faced by children within particular age groups when exposed to PC may vary quite significantly and some children will be more vulnerable than others, such as neurodiverse children and children whose gender, race, and sexuality may impact the harm they experience from content (See Draft Children's Register of Risks for more information, Volume 3, Section 7). Therefore, while there may be some unintended adverse impacts on some children who would be less severely affected if exposed to such content, this may not be the case for all children across a particular age group for whom this additional protection may provide significant benefits. We may evolve this approach over time as we develop more specific evidence on the nature of these harms.

20.123 In reaching this view, we have sought to strike a proportionate balance between ensuring that children who are more vulnerable to harm from encountering PC are provided with an adequate degree of protection, in ensuring they are less likely to be exposed to such content repeatedly and in large volumes, while also not restricting children who are less likely to require such protection from seeing any PC at all in their recommended feeds. For these reasons, and the reasons set out in relation to Measure RS1 as applicable here, we consider that to the extent that this measure interferes with service providers' or users' rights to freedom of expression, those impacts are limited and no further than needed to secure adequate protections for children. We therefore consider the degree of interference with users' and service providers' rights to freedom of expression is proportionate, particularly in light of the benefits to children this measure will help to secure.

20.124 We also consider the impacts on users' rights to privacy will be very similar to those in relation to Measure RS1 above and have not identified any additional privacy or data protection impacts relating specifically to this measure. Therefore, for the reasons set out in relation to Measure RS1 above, as also applicable to this Measure (RS2), we do not consider there to be any material impact on user's right to privacy. We consider the measure to be proportionate on this basis, particularly considering the benefits to children that it would secure.

20.125 As explained in the Age Assurance, Section 15 within this Volume, we are recommending that service providers in scope of this proposed measure also use highly effective age assurance to identify child users who should benefit from the protections offered by this proposed measure. Refer to Measure AA5 in the Age Assurance section within this volume for more information. We discuss the rights impacts we expect to arise in relation to the use of age assurance in that section and we do not consider them separately here.

Impacts on services

20.126 We consider separately below the direct costs of modifying the service to implement the proposed measure, costs related to age assurance, and the potential for an indirect cost to services resulting from lost revenue.

20.127 Table 20.3 below presents quantified estimates of direct cost estimates, based on the assumptions summarised in this sub-section. Although we have drawn on available evidence and expert input, our quantitative estimates of costs should be interpreted as indicative. Real world costs will depend on the specific recommender systems and associated systems

used by services. Table 20.4 below presents illustrative costs of age checks which is discussed in more detail in Section 15, Age Assurance.

Table 20.3: Summary of direct cost estimates

Activity	One-off implementation cost	Ongoing annual cost
Implementing the measure	£18,000 to £89,000	£4,000 to £22,000
Linking Highly Effective Age Assurance to the measure (if not in scope of RS1)	£9,000 to £36,000	£2,000 to £9,000

Source: Ofcom analysis

Table 20.4: Illustrative cost estimates of age checks via third-party age assurance providers[787]

Service size	Existing UK user base	New users each year	Age assurance for existing users	Age assurance for new users (annual ongoing cost)
Smaller service	100,000	10,000	£5,000 to £20,000	£1,000 to£2,000
Larger service	7,000,000	70,000	£350,000 to £1,400,000	£4,000 to £14,000

Source: Ofcom analysis. Note that the above estimates are based on age checks being conducted for all users, which is likely to be an upper bound and may overestimate costs, as we explain below under 'Costs related to age assurance'.

Direct costs of implementing the measure

20.128 We understand that there will be costs associated with adjusting a recommender system to reduce the prominence of content that is likely to be PC on the recommender feeds of children. The sources of costs are similar to those associated with Measure RS1, outlined above. As discussed there, we understand that service providers that already use content recommender systems are likely to have engineering skill and computing infrastructure but would still expect a cost to be incurred to adjust the recommender to ensure this content is reduced for children. We set out below our understanding of the activities and associated costs on providers of steps described in the explanation of the measure above that may need to be implemented to follow this measure.

20.129 We understand that the first two sets of activities involved in implementing this measure are likely to be similar to those for Measure RS1, whereas the third step differs. Service providers will incur costs in relation to:

1. Using relevant available information to identify content likely to be PC.
2. **Making the signal available to the recommender system**. We set out in previous sub-section 'Explanation of the measure' for RS1 the considerations of how a service may implement this measure. Everything described in that sub-section for PPC is relevant to this measure, but for PC; and

[787] For further detail on age assurance cost analysis and estimates for more sizes of services see Annex 12

3. Modify the recommender system to significantly reduce the prominence of content likely to be PC in children's recommended content. Service providers would need to ensure that safety information is appropriately actioned by the recommender system. This could involve re-programming the component of the recommender system that is relevant for re-ranking to downrank content likely to be PC from children's recommended content. This might look like redesigning the re-ranking algorithm so that it can assign a lower relevancy score for content likely to be PC.

20.130 We understand that the main cost in implementing this measure would be labour input from software engineers, machine learning teams, and data specialists. Service providers will likely need to test their recommender system once it has been redesigned, and additional time may be required to conduct these to ensure effectiveness.

20.131 We estimate that for a service provider to undertake the three activities above could require a one-off direct build effort of approximately 8-20 weeks of labour time split across roles including software engineers, machine learning engineers and data scientists. We have assumed that this time is matched with an equal amount of non-software engineering time (e.g. project management, legal, trust and safety). This range of costs is broadly similar but slightly higher than our estimated time for Measure RS1 to account for some potential extra time needed to implement the process to significantly limit prominence of content likely to be PC. This may have a somewhat higher degree of complexity than filtering content outright (as recommended under Measure RS1), given other factors which might impact prominence (e.g. virality). Using our assumptions on labour costs required for software engineering work set out in Annex 12, we estimate one-off direct costs in the region of £18,000 to £89,000.

20.132 The cost of this measure for a given service provider will be impacted by the existing design of the recommender system. We consider that costs will be higher for providers with more recommender systems operating, and where systems are complex (for instance serving more users in more languages) but do not already have a mechanism for limiting the prominence of certain types of content. As set out in the 'Effectiveness' sub-section above, some large services have already designed their recommender system to ensure that certain types of content are downranked. It may be more straightforward for service providers to implement this measure if they already have the infrastructure in place to reduce the prominence of specific types of recommended content and only require more minor adjustments to implement this measure.

20.133 As noted in Measure RS1 above in the 'Direct costs of implementation' sub-section, there is potential for a service provider to incur model training costs depending on how the provider chooses to implement this measure, and that these could be material. We consider that there may also be additional business oversight and coordination costs associated with changing products. Larger businesses may use more complex processes for system changes and face significant review, communication and legal processes to implement changes to their services. We would expect the oversight and coordination costs to be largely correlated with the size of the company, but do not have sufficient information to be able to quantify these.

20.134 We would also expect a service provider to incur ongoing costs, including maintenance to ensure that it continues to function as intended. There may also be ongoing costs of an extended product management cycle where service providers may have additional objects to consider as part of ongoing management, for instance where there are additional

variables to observe in terms of how the measure is performing. In line with our standard cost assumptions set out in Annex 12, we assume this to be approximately 25% of the initial set-up costs, ranging from approximately £4,000 to £22,000 per year.

20.135 As described above on Measure RS1 (see Direct costs of implementation sub-section), this is likely to involve similar activities and teams to this measure (Measure RS2). We believe that the implementation of Measures RS1 and R2 would likely be undertaken jointly where a service provider is in scope of both measures due to risks across PPC and PC content, and as a result there could be some synergies when providers are making these changes simultaneously. While these synergies could be significant in the case of some service providers, there is a high degree of uncertainty about the degree and variation of the overlap of costs of the two measures between services. Therefore, we have not quantified an estimated cost reduction for service providers implementing these measures simultaneously.

Costs related to age assurance

20.136 Service providers in scope of this measure may apply highly effective age assurance to target this measure at children, although this is not specifically recommended. The costs of highly effective age assurance are covered in the Age Assurance, Section 15 within this Volume, and our discussion of Measures AA5 and AA6.

20.137 The same cost implications of age assurance and linking age assurance to the recommender system discussed above in relation to Measure RS1 in the 'Impacts on services' sub-section for that measure would apply here in relation to RS2. However, costs would only need to be incurred once if a service provider is in scope of more than one of the proposed Recommender System measures.

Indirect costs to services resulting from lost revenue

20.138 We consider that there is the potential for this measure to have similar indirect costs on service providers as Measure RS1, as set out in the Indirect costs to services sub-section for RS1. Similar to that measure, where the prominence of likely PC is significantly reduced for children that would have been more prominently recommended otherwise, we believe any business impacts from this are justified. We consider there is likely to be some non-harmful content which is reduced in prominence for children due to this measure, with some potential business impact, but we believe this is limited as service providers can recommend other non PPC/PC content instead. We also believe that the countervailing indirect reputational and engagement benefit to service providers which are associated with less harmful content to children described in Measure RS1 also applies here.

Which providers we propose should implement this measure

20.139 We propose to recommend this measure for all U2U services likely to be accessed by children that have a content recommender system (as set out in 'An explainer: What is a recommender system?') and are medium or high risk for at least one kind of PC (excluding bullying content). We are minded to extend this measure to also include service providers with medium or high risk of certain categories of non-designated content, namely body image content and depressive content, subject to the outcome of the consultation on our proposals to classify body image and depressive content as non-designated content. Should new evidence emerge of other forms of NDC, that are also likely to cause cumulative harm, we will seek to amend these measures accordingly.

20.140 As explained in Section 15, Age Assurance, within this Volume, we consider that service providers in scope of this measure should apply highly effective age assurance to target the measure at children and achieve the intended effect – see proposed Measure AA6. In assessing the proportionality of this measure for different kinds of providers, we therefore consider the impacts of both this measure and related age assurance measure in the round.

20.141 Recommender systems are a key pathway for children to encounter PC (aside from bullying). They can introduce children to this content for the first time and facilitate repeated engagement, leading to cumulative harm. This can occur even on services where PC is prohibited, given the challenges of moderating content at scale resulting in this content often appearing in recommender feeds. The proposed measure recommends service providers take a precautionary approach to significantly limit the prominence of recommended content for children, when they have sufficient indications that it may be harmful to them, even when this is not yet confirmed. We expect this can make a significant contribution to protecting children from encountering PC.

20.142 We consider that this measure should apply to service providers whose risk assessment indicates that children face a medium or high risk for at least one kind of PC, or of the types of NDC we are minded to include, subject to consultation, because this is where this measure will create material benefits by helping prevent children encountering this content.

20.143 The estimated costs of this measure for service providers can be significant. We recognise the possibility that a minority of small businesses in scope of this measure could struggle to carry this cost. Providers may be discouraged from offering recommender feeds, and where recommender systems are integral to business models, it could discourage some service providers from serving UK users. This could harm users who benefit from accessing these functionalities and even services.

20.144 However, in most cases we consider that the costs to service providers will vary depending on the complexity of each provider's recommender system. Costs will be higher for service providers with more complex recommender systems, including where these cater for large volumes of users and multiple languages, and we believe these providers will typically have greater capacity to implement changes. We have designed this measure to allow some flexibility in how it is implemented, enabling service providers to manage their costs accordingly. We also believe that services running recommender systems have the necessary technical capabilities to implement this measure. Our assessment of costs and judgement of proportionality is based upon considering that service providers with these systems already have a level of technical maturity to allow their recommender systems to receive and action relevant information.

20.145 The related age assurance costs are expected to largely scale with size of service provider and may also scale with the level of risk. The riskiest services are more likely to limit the prominence of large volumes of content likely to be PC from their recommender feeds, which may motivate a greater proportion of adult users to conduct an age check and have this content more prominently in their feeds. While costs may be higher for such service providers, the benefits to children's safety from this measure is also higher on the riskiest services.

20.146 Overall, we recognise the measure imposes material costs that could lead some loss of choice if small services struggle to shoulder the burden of this measure. We nonetheless consider it proportionate to apply this measure to all services who have a recommender system and are medium or high risk for at least one kind of relevant risk (regardless of size)

given our view of the effectiveness of the measure and of the important role played by these systems in exposing children to harm related to PC.

20.147 We do not recommend this measure for providers where the risk of PC is low, because the measure would then have limited benefits for children's safety, if any, while its impacts on service providers and adult users would still be material.

20.148 We have provisionally concluded to recommend this measure to all U2U services likely to be accessed by children (regardless of size) that have a content recommender system, and that are medium or high risk for at least one kind of PC (excluding bullying), and are minded to also apply to services that are medium or high risk for at least one of body image and depressive content NDC, subject to consultation on these harms as described above.

Provisional conclusion

20.149 Given the harms this measure seeks to mitigate in respect of all kinds of PC excluding bullying content, as well as the risks of cumulative harm that recommender systems pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU F2 in Annex A7.

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

Explanation of the measure

20.150 The proposed measure should provide children a means of expressing negative sentiment towards content that they encounter that is harmful to them.[788] This could include content that they find distressing or upsetting. This negative sentiment should result in negative feedback directly into the recommender system, so that content similar to this is limited in prominence and, therefore, appears less frequently for that user in the future. The visual appearance of this measure may vary, and we will not be prescribing how this measure should be visually designed and integrated into a user interface. However, the proposed measure should result in a clear pathway for children to provide negative feedback into the recommender system when they encounter content that is harmful to them.

20.151 In terms of the user-journey of this measure, services may wish to seek more granular user feedback on content. For example, this may take the form of a functionality that allows

[788] We expect services to take appropriate action against bullying content in relation to CM1 and RS3, however given limited evidence of the connection between bullying content and recommender systems a service provider would not be in scope of the measure if the children's risk assessment does not identify any risk of PC other than bullying on the service and there is no other risk of harm. If we receive more evidence of the connection between bullying and recommender systems, we may be able to revise this position. We are also minded to extend this measure to certain categories of non-designated content, namely body image content and depressive content. However, this would be subject to the outcome of the consultation on our proposals to classify body image and depressive content as non-designated content. Based on what we know about these kinds of proposed NDC, it is highly likely this will occur on recommender systems. We therefore currently have evidence of all harms occurring on recommender systems except bullying. This may be due to the fact that bullying content is associated with behaviour between users rather than content.

children to provide a reason for expressing a negative sentiment, so that they can reduce the prominence of similar content depending on the reason provided. One way of achieving this could be to apply an appropriate downranking to this content by way of feedback to the recommender once the child has expressed negative sentiment. For example, where a reason is provided, a service may decide to downrank distressing content with more severity than content that was shocking but not necessarily harmful. Alternatively, services may decide not to prompt children to provide a follow up reason for expressing negative sentiment and choose to enable negative feedback to the recommender on all content that children express negative sentiment towards, regardless of their reason for doing so.

20.152 When children express negative sentiment on an individual piece of content, this should result in **similar content** also being limited in prominence, rather than needing to repeat this action in future on the same type of content.

20.153 By similar content, we refer to any content that shares significant characteristics with a given piece of content (i.e., content a user has expressed negative sentiment on). Significant characteristics may include, but are not limited to:

- Subject matter: the topics, themes, or issues addressed in the content (e.g., weight loss and dieting);
- Metadata: information about the content such as a tags, hashtags, categories, and keywords associated with that content.

20.154 While we will not be specifying how services determine the similarity of content, we would expect them to use any relevant available information that is considered indicative of significant characteristics to determine and infer content similarity and limit the prominence of such content in a manner they consider proportionate. As noted below, there are several ways a recommender system received explicit and implicit user feedback from children. Often this feedback is used to assume positive engagement, even when this is unintentional. This measure counteracts the range of inferred positive feedback generated from children's engagement with recommended content.

20.155 This measure differs from the Measures proposed in the User reporting and complaints (Section 18). Unlike reporting, this measure aims to create a system whereby users' negative feedback of content directly leads to content similar to this being given less prominence in future recommendations to the same user, and only in their feed. The proposed measure will not necessarily result in the content that the child has expressed negative sentiment towards and similar content being less prominent for all children.

20.156 In contrast, the reporting and complaints mechanisms alert services to content likely to be harmful to children, often inputting into content moderation systems by flagging this for review. Services may then take time to consider a user report and content may ultimately be removed off the back of the original report. The provision of negative feedback on recommended content would not lead to removal of that content. Instead, it directly informs the recommender in real time to reduce the visibility of this content for the individual user in a way that reporting unlikely to be able to do so as quickly.[789] [790]

[789] Wang, Y et al., 2023. Learning from Negative User Feedback and Measuring Responsiveness for Sequential Recommenders. [accessed 22 April 2024].
[790] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.

20.157 In addition, the proposed Measure US4 in Section 18 of this Volume may be applicable in conjunction to this measure as US4 provides children with additional information each time they take restrictive action against a piece of content.

Definition Box 4: What do we mean by user feedback?

By **user feedback,** we mean the various types of data that helps the recommender systems learn about users' preferences, behaviour, and interactions with content. There are two broad types of user feedback:

- **Explicit feedback:** this refers to direct and intentional actions taken by users to express their preferences and sentiment on content, for example likes/dislikes. Though it can it vary across services; explicit feedback provides recommender systems with clear and unambiguous information about a user's preferences. Depending on the service, reporting/complaints can also be forms of explicit negative feedback.
- **Implicit feedback**: this refers to feedback into the recommender systems that the user may not have intended. Implicit feedback can involve the number of times a user clicks on an item, the amount of time they spend interacting with it (e.g., watch time), and how they scroll through content.

This measure focuses on the use of a dismissal functional that allows users to give explicit negative feedback to ensure children receive fewer recommendations of content they do not want to see and find distressing, helping to shape their own recommender feed in ways that are safer and more relevant and reducing the risk of exposure to harmful content and cumulative harm.

20.158 This measure has the potential to help manage the risk of children being impacted by harmful content, including NDC. Children will be able to express negative sentiment towards NDC which is key in cases where children are repeatedly exposed to content they may find distressing. The negative feedback to the recommender should result in this being appropriately limited in prominence for the child expressing negative sentiment so that this has the effect they are less likely to see this content in future. This signal may also provide relevant information to the service provider about which content could be NDC. An additional benefit to this measure is to support the correct implementation of Measures RS1 and RS2. Where PPC or PC content has not been sufficiently filtered out or limited in prominence respectively under Measures RS1 and RS2, this measure can act as a safety net. This measure therefore applies to all content, but as described below, service providers that want to vary the weight (i.e. the impact) of a child's negative feedback on content may want to apply neutral weightings to content out of scope of the Act.

20.159 Without a function or mechanism to explicitly signal negative feedback, children's expressions of distress or negativity towards harmful content, such as disapproving comments or hovering over content in distress or disbelief can be misinterpreted as positive engagement by services. This misinterpretation may inadvertently prompt the recommender system to serve more harmful content. Service providers may decide to apply RS3 to all users. Alternatively, service providers may limit this measure to child users in cases where they apply highly effective age assurance under AA5 and AA6 (Section 15 within this Volume). In this case, the proposed measure would be made available on the recommender feeds of accounts held by children and the signal given by the child when providing negative feedback should have the outcome that the child see's less of this content. Where the

feature is utilised, services may then prompt the user to select a reason or description of what the user experienced.

20.160 Alongside explicit user signals such as likes, comments, and upvotes, we understand that recommender systems use a variety of implicit user signals as inputs to understand user preferences, such as watch time.[791] Children may not always be aware, particularly young children, that a recommender system exists and is collecting a variety of data about their viewing habits to make inferences about their preferences. For example, when a user engages with content by leaving negative comments, the act of commenting may be taken as a positive feedback signal for the recommender system.

20.161 Services that use recommender systems tend to make it easier for users to express positive sentiment on content (e.g. with a like button) than they do negative sentiment., Recommender systems can interpret, albeit unintentionally, unwitting engagement (e.g., watch time, hovering over content) on harmful content as positive engagement. To manage this risk, we consider it particularly important to allow children to dismiss content they find distressing.[792]

20.162 In our research with young people and professionals exploring experiences of suicide, self-harm and eating disorders, participants provided suggestions to improve the safety of children online in relation to these harms. These included, providing clear and pro-active ways for users to tailor the kinds of content they receive via content recommender feeds.[793] Children may also encounter PC via recommender systems, in particular content depicting or encouraging violence. In our research with children, they told us that they felt they had no control over the content they see via recommender feeds.[794] Content such as ordinary dieting content or content focused on fitness may not be considered likely to be PPC or PC but, when encountered alongside NDC, may be cumulatively harmful for individuals with existing vulnerabilities. Although Measures RS1 and RS2, if implemented effectively, should mean this content is filtered from the recommender feeds of children, this measure (RS3) provides an additional layer of protection from such content. This measure enables children to say when they want to see less of specific kinds of content.

20.163 There is some evidence suggesting that child users would materially benefit from features that can mitigate the risk of children's exposure to harmful content[795] via recommender systems by designing the system to respond to negative sentiment in real-time and adjust content feeds accordingly. In addition, our report into evaluating recommender systems explained how these systems are nuanced and can be designed to ensure that the extent of reducing the prominence of certain categories of content (or content containing certain characteristics) is proportionate to the level of risk of that content.[796]

---

[791] Wall Street Journal, 2021. Investigation: How TikTok's Algorithm Figures Out Your Deepest Desires. WSJ, 21 July. [accessed 24 April 2024].
[792] Bengani, P., 2022. What's right and what's wrong with optimizing for engagement. Medium, 27 April. [accessed 24 April 2024].
[793] Ofcom, 2024. Online Content: Qualitative Research, Experiences of children encountering online content promoting eating disorders, self-harm and suicide.
[794] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.
[795] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.
[796] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content.

Definition Box 5: Defining negative sentiment

By **negative sentiment**, we mean the unfavourable or adverse emotions, or feelings experienced by children when encountering harmful content. This can include anxiety, sadness, anger, fear, frustration, or any form of distress. In the context of children encountering harmful content, a child user may not always be able to recognise, understand or express distress in a constructive way. It is important for online services to consider this risk when designing their recommender systems and user interaction features.

20.164 To ensure children can explicitly express negative sentiment which will result in negative feedback on content harmful to them, we recommend that services implement the following changes to content feeds underpinned by recommender systems.

20.165 On children's accounts, there should be a feature that allows the user to privately express negative sentiment on content encountered via recommender feeds. The exact form and design of this feature can be decided at the services discretion. For example, a "show me less of this" button or "I don't want to see this" are possible options for this.

20.166 Service providers may prompt the user to provide a follow-up reason for why they expressed negative sentiment. These can take the form of follow-up reasons that should sufficiently reflect feelings of distress. Prompting users to provide a follow up reason as to why they are expressing negative sentiment could be a helpful step that can help services gain insight into what type of content children are expressing negative sentiment and identify potential cases of NDC. For example, where a large number of children express negative sentiment on a particular piece or category of content, services can become aware of NDC, PC, and PPC.

20.167 Service providers may choose to prompt users about why they are expressing negative sentiment, because this may enable the action to be more targeted. For example, depending on the follow-up reason selected, the resulting signal into the recommender system can be proportionately weighted to determine the extent of reduction of prominence for that content. We will not prescribe a weighting scheme for this measure, but the outcome should be that content where children have expressed negative sentiment should be significantly limited in prominence for that child.[797] In the absence of a prompt for why children are expressing negative sentiment, it may be appropriate for service providers to reduce the prominence of all content where children express negative sentiment.

20.168 However, when considering whether a service wishes to prompt for a reason, services should also consider the risk of notification 'fatigue'; where too many notifications become ignored due to consuming excessive time and mental energy (particularly when not perceived to be serious or relevant).[798]

Managing the risk of misuse and undue suppression of content

20.169 We have considered the risks associated with the misuse of features that allows users to provide negative feedback into recommender systems. However, we understand that this risk can be sufficiently managed by ensuring such a feature is designed with the following guardrails:

[797] Weighting scheme refers to the various degrees of influence/significance different factors are given within the recommender system.

[798] Cash, J., Pharm, B. and Pharm, D., 2009. Alert fatigue, *American Journal of Health-System Pharmacy*, 66 (23). [accessed 24 April 2024].

a) **The signalling of negative feedback is private:** this means that the number or volume of users signalling negative sentiment is not visible publicly nor to the content creator. Publicly displaying the volume of negative sentiment (e.g. dislikes) can encourage the adversarial use of recommender systems to artificially suppress content. For example, users can carry out coordinated "dislike" campaigns on other users to deliberately suppress recommendations.[799] Our measure designs out misuse as it is private in that the number of users that have dismissed the content will not be visible and the effect of negative feedback is isolated to the user giving the feedback.
b) The **feature is only available on content recommended feeds:** we are only recommending this measure for content that is encountered directly by a recommender system, not indirectly via other functionalities such as searching for the content, direct messaging, or via a chronological feed.

20.170 The measure may, in practice, result in children using the feature on content they simply lack interest in due to preference (e.g., not liking a certain football team). Where service providers choose to ask for a reason as to why a child does not want to see content, it may be possible to enable children to signal negative feedback towards PC, PPC or NDC and therefore limit the prominence of this content according to the child's sentiment. In practical terms, this may mean assigning varying levels of importance to negative feedback signals to moderate how prominent this content is on the feed of the child expressing negative sentiment.[800]

20.171 However, where services do not ask the user to provide a follow up reason, services should ensure that all content that is similar to which children express negative sentiment is limited in prominence so that the child sees less of this both then and in the future. As currently proposed, this measure does not require service providers to determine why children are expressing negative sentiment.

Effectiveness at addressing risks to children

20.172 One of the ways of managing the risk of a recommender system amplifying harmful content to children is through allowing them a means of signalling negative sentiment on content that might be harmful to them.

20.173 An industry paper by senior software and machine learning engineers at Google, Meta, Snap, and OpenAI published in 2023 highlighted the importance of negative user feedback into recommender systems in empowering users to shape their content feeds. Published at the ACM RecSys 2023 conference, we believe this research is particularly relevant because it acknowledges the importance of designing recommender systems that are more responsive to explicit negative user feedback.[801] We acknowledge that deploying models that can learn from negative user feedback is a notorious engineering challenge and a live area of experimentation and testing. However, we also acknowledge that this is an area of innovation as online services aim to make recommender systems more relevant and safer

[799] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content
[800] By assigning different levels of importance to certain signals, the recommender system can distinguish between different types of user feedback and reduce the prevalence of that content in accordance with the strength of the signal.
[801] Global conference on recommender systems organised by the Association for Computing Machinery (ACM). See: ACM (2024), 17th ACM Conference on Recommender Systems. [accessed 24 April 2024].

for their users.[802] We consider this measure consistent with emerging design practices of content recommender systems.

20.174 In addition, there is a growing body of research that highlight the importance of user controls and managing the prevalence of harmful content on recommender feeds. This includes behavioural and qualitative evidence of children expressing the desire for greater control over algorithmically curated feeds. This evidence includes:

i) Ofcom's Children's Media Lives 2023 qualitative study found that children's viewing of content was mostly passive and that most of the content they saw was served to them via recommender systems rather than actively searched for. Several of the children in the study reported seeing content appear in their feeds that they had not sought out and sometimes that they would rather they had not seen.[803]

ii) Ofcom's report into the experiences of children encountering online content relating to eating disorders, self-harm, and suicide found recommender systems to be the pathway in which they first encountered content relating to eating disorders, self-harm, and suicide.[804] In another study, Children and young people also told us that they use features to minimise the risk of seeing similar content in the future.[805] Ofcom found that many children expressed that they feel they have no control over the content recommended to them by recommender systems, particularly violent content.[806]

20.175 In the absence of a dismissal function that allows users to signal negative sentiment and steer the recommender system away from suggesting harmful content, children may or may not engage with harmful content recommended to them. When services use watch time as a positive engagement signal, children are at risk of unwittingly engaging with content (e.g., simply by watching it). Definition Box 4 sets out how implicit feedback can provide recommender systems with data about users' behaviour. While some child users might not explicitly or intentionally signal their preferences, their viewing patterns, search history, number of clicks, and watch time on certain content still serve as learning data for the recommender system from which it can make inferences about the child's preferences.

20.176 It is generally accepted across the computer science community and U2U services that, alongside user reporting, negative sentiment resulting in negative feedback relating to content can provide an excellent signal of harmful content when this is provided by a diverse range of users.[807] Recent research has focused on utilising negative feedback mechanisms into recommender systems to reduce unwanted content recommendations, When applied to content harmful to children, the paper demonstrates that recommender systems can be responsive to negative sentiment and use negative feedback to reduce the risk of adult and

---

[802] Wang, Y et al, 2023. Learning from Negative User Feedback and Measuring Responsiveness for Sequential Recommenders. Published by the ACM for the RecSys 2023 Conference (Industry Track) held in Singapore. [accessed 24 April 2024].

[803] Ofcom, 2023. Children's Media Lives.

[804] Ofcom, 2024. Online Content: Qualitative Research, Experiences of children encountering online content promoting eating disorders, self-harm and suicide.

[805] Adults in the sample were reflecting back to their experiences during childhood. Ofcom, 2024. Online Content: Qualitative Research, Experiences of children encountering online content promoting eating disorders, self-harm and suicide.

[806] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children.

[807] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content

child users encountering harmful content to them.[808] There is also a growing body of evidence that demonstrates why this measure would be an effective tool for children. User feedback into recommender systems minimises the risk of users encountering low-quality and harmful content and services get increasingly more nuanced data on what types of content users want to see less of.[809] This measure (RS3) would help services ensure that children who find themselves being recommended streams of harmful content have a means of giving immediate feedback to the recommender system to express that they want to see less of this content. We therefore consider this an important measure for improving user safety.

20.177 Our rationale for focusing on providing users with a means of expressing negative sentiment to form negative feedback signal directly to the recommender is as follows:

20.178 Positive feedback bias occurs when recommender systems, by design, have more exposure to positive engagement signals relative to negative signals that indicate negative sentiment. To counter this, some services have introduced means for users to provide negative feedback. For example, a Reddit community-based voting system allows users to downvote content, and the recommender system is designed to minimise the visibility of downvoted content.[810] This can help filter out low-quality content and promote high-quality content to the top of the feed. If a user hides a lot of posts about a particular topic, the recommender system will be less likely to recommend posts about that topic to that user in the future.[811]

20.179 There is a growing effort from services to incorporate negative feedback into recommender systems as a means of promoting user safety. For example, in their Engineering Blog, Pinterest recognised the importance and value of collecting negative user feedback to ensure the recommender system is optimised for what the user does not want to see.[812]

20.180 Allowing users to signal negative sentiment on content resulting in a negative feedback signal is common across U2U services, and there is evidence of several U2U services providing users with a means of signalling disapproval on content recommended to them. Such features typically take the form of feed controls such as muting or hiding certain topics, blocking certain keywords, or hiding certain pages.

20.181 Based on current industry practices, several services acknowledge that users should have some means of providing negative feedback into their recommender systems. Some examples of current practices include:

a) In August 2022, Meta announced that they are testing new ways of controlling what content users see on the Instagram recommender feed.[813] As part of these trials, users have been given three new feed controls. This includes a "not interested" control that removed the post from the user's feed immediately while aiming to suggest fewer

---

[808] Wang et al., 2023. Learning from Negative User Feedback and Measuring Responsiveness for Sequential Recommenders. Published by the ACM for the RecSys 2023 Conference (Industry Track) held in Singapore. [accessed 24 April 2024].
[809] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content; Reddit, 2023. How does voting work on Reddit [accessed 26 April2024]; New America, Everything in Moderation, 2019. Case Study: Reddit [accessed 24 April 2024].
[811] Reddit, 2024. Reddit's Approach to Content Recommendations. [accessed 24 April 2024].
[812] Pinterest Engineering, 2022. How Pinterest Leverages Realtime User Actions in Recommendation to Boost Homefeed Engagement Volume. [accessed 24 April 2024].
[813] Meta, 2022. Testing More Ways to Control What You See on Instagram. Meta Newsroom, 30 August. [accessed 24 April 2024].

similar posts. There is also a snooze feature that allows users to take a break from certain posts for 30 days. The third feature allows users to adjust the level of sensitive content [814] the user wish to see. Users can choose whether they want to see more, less, or a 'standard' amount of sensitive content. This test by Meta indicates that sensitive content detection and labelling is technically feasible, and varying weights can be applied to different user controls. Additionally, Meta introduced feed controls on Facebook in 2022, enabling users to see less of what they don't want to see.[815] While these are user optimisation tools, they highlight that services consider is important to allow users some degree of control over recommender feeds.[816]

b) In January 2021, LinkedIn introduced new controls to help surface content that is relevant to users, and this includes the option to say, "I don't want to see this". In May 2022, LinkedIn built on this feature and introduced more options for users to signal why they might not be interested in the content. A valid reason for this might be that content is unprofessional and violates LinkedIn's professional community policies (for example, if the content is highly political or controversial).[817] If a user utilises this feature on content, it is processed as negative feedback by their recommender system, and LinkedIn will endeavour to show less of that content to that user. While the LinkedIn practices relate to content that might be considered unprofessional for some users, the professional community polices include categories of content that include priority content. This indicates that is technically feasible to design an intervention that allows users to signal negative sentiment on certain types of content.

c) In June 2022, X (formerly Twitter) introduced a "downvote" button. Unlike the ''like'' button, downvotes cannot be seen by the original poster or other users – it is therefore a private means of signalling negative sentiment. However, this feature is only available on replies to original posts. X has stated that this button is used to signal offensive content and prioritise high-quality content. In addition to downvoting, X allows users to mute or block users and topics (such as keywords and hashtags).[818] As part of their "quality and safety ranking", X state that blocking or muting an account is taken as a strong signal that a user does not want to see those posts.[819] As part of their recommender system overview, X sets out the various components/modules that make up the system, and the signals that inform recommendations. In this overview, it shows how negative sentiment is used to inform the underlying machine learning system.

20.182 In summary, large U2U services have deployed and tested features that allow users to signal negative sentiment on recommended content, and to ensure that this is processed as negative feedback by the recommender system. While the industry practices presented above are typically considered user optimisation tools, we consider these to sufficiently indicate that feed controls for the purposes of reducing the prevalence of content harmful to children are technologically feasible and is a proportionate measure for children.

---

[814] Instagram (Meta), 2022. Updates to Sensitive Content Control. [accessed 24 April 2024].
[815] Meta, 2022. New ways to customize your Facebook feed. [accessed 24 April 2024].
[816] Instagram (Meta) (Mosseri, A), 2023. Instagram Ranking Explained; Meta, 2019 Using Surveys to Make News Feed More Personal. [accessed 24 April 2024].
[817] LinkedIn, 2023. Professional Community Policies. [accessed 24 April 2024].
[818] X, (no date). How to use advanced muting options. X Help Center. [accessed 19 April 2024].
[819] X (no date). About specific instances when a post's reach may be limited. X Help Center. [accessed 19 April 2024].

Rights assessment

20.183 This proposed measure recommends services provide children a means of expressing negative feedback towards content they encounter that is harmful to them. We expect this measure will significantly reduce children's exposure to harmful content. This aligns with the legitimate aims of the Act to secure a higher level of protection for children than for adults and which imposes duties on services which requires them to use proportionate measures to protect children from content that is harmful to them.

20.184 In implementing this measure, there is a potential impact on users' rights, in particular, their rights to privacy (Article 8 of the ECHR) and to freedom of expression (Article 10 of the ECHR). We have considered the extent to which the degree of interference with these rights is proportionate and in doing so, our starting point is to recognise that the Act requires services to take proportionate systems and processes to protect children from encountering content that is harmful to them and to mitigate any impact of harm presented by content that is harmful to them. We therefore consider that a substantial public interest exists in measures which aim to protect children from content that is harmful to them.

Freedom of expression and association

20.185 We are only recommending this measure for content that is encountered directly by a recommender feed, not indirectly via other functionalities such as searching for the content, direct messaging, or via a chronological feed. This measure does not involve services taking steps in relation to all users but rather recommends that the effect of the negative feedback is isolated to the user giving feedback. Recommender systems can individualise feedback to accommodate the diversity of user preferences, meaning that the effect on negative feedback on future recommendations can be limited to only those signalling negative feedback.

20.186 As a result, we consider that any impact on users' rights will be limited as only the user in question who has expressed negative feedback towards a particular piece of content should see that content and similar content is limited in prominence on their recommended feeds, and where they express this preference they are exercising a choice to restrict the information they choose to receive, which is particularly important where this can help to protect child users from harm.

20.187 Given the content on which negative feedback has been signalled towards will still be available and accessible on the service by other users, including in their recommended feeds, we consider that services can implement this measure without creating any material adverse effect on other users' rights to freedom of expression, and to the extent it may have an impact on their own rights to freedom of expression, we consider this to be very limited and justified.

20.188 In addition, we consider that this measure could have positive impacts on children's rights to freedom of expression and freedom of association as we expect it will result in children having options to limit their exposure to content which would be harmful to them (particularly NDC), which could result in safer spaces online where children may feel more able to join online communities and receive and impart (non-harmful) ideas and information with other users, providing significant benefits to children.

20.189 While we acknowledge that some services might choose to extend this functionality to all users, rather than targeting it only at children, if they choose to do so we do not consider this would lead to any material adverse impacts on freedom of expression for the reasons

explained above. Indeed, while not the specific aim of this measure, this could lead to benefits to adult users if they have additional options to protect themselves from exposure to harmful content too.

Privacy

20.190 We also consider the impacts on users' rights to privacy will be very similar to those in relation to Measures RS1 and RS2 above. In particular, we are not recommending that services process or retain any additional personal data, and would not expect them to need to do so (or at least to any material extent) beyond what they would already be processing or retaining in order to provide their users with personalised recommended feeds. We also note that to the extent they do process any personal data in implementing this measure they would need to do so in accordance with data protection legislation requirements. Therefore, for the reasons set out in relation to Measure RS1 above, as also applicable to this Measure RS3, we do not consider there to be any material impact on user's right to privacy and the measure to be proportionate on that basis, particularly considering the benefits to children that it would secure.

Impacts on services

20.191 We consider separately below the direct costs of modifying the service to implement the proposed measure, costs related to age assurance, and the potential for an indirect cost to services resulting from lost revenue.

Direct costs of implementing the measure

20.192 We describe our quantified estimates of direct costs, based on the assumptions described in this sub-section. Although we have drawn on available evidence and expert input, our quantitative estimates of costs should be interpreted as indicative. Real world costs will depend on the specific recommender systems and associated systems used by services.

20.193 There are likely to be significant costs associated with introducing a user feature to allow users to privately express negative sentiment and adjusting the recommender system to process negative user feedback in a way that can reduce prevalence of specific categories of content on the recommender feeds of individuals. Doing so would require sophisticated computing infrastructure, specialist engineers, and ongoing maintenance. We set out the direct costs we would expect these services to incur when deploying this measure where services do not already have such features. However, more sophisticated services may already have elements of the measure in place. As described above, several services have begun to introduce user control tools that may be relevant to this measure, and therefore the incremental cost to implement the proposed measure would be lower for such services than our estimates set out here.

20.194 **Designing and testing a feature that allows the user to privately express negative sentiment**. Setting up the ability for children to give negative feedback on recommended content would require services to introduce a new user-facing element on recommendation feeds, likely involving input from front end engineers to design, develop, and deploy. Service providers may prompt the user to provide a follow-up reason for expressing negative sentiment, and this will involve additional costs. User testing may be required to and test different design options, for instance where to position the button and what it should be called, which may involve A/B testing for a service to be able to conduct user experimentation as part of optimising. Services may want to carry out user research into what response categories are most appropriate and used as intended by children.

20.195 **Adapting the recommender system to process negative signals**. Services would need to modify the recommender system so that it can receive negative user feedback on content and adjusting future recommendations for similar content accordingly. This would require labour time to implement from data engineers and data scientists. We understand that this step is technically feasible, though can be challenging. Service providers can use information as appropriate to determine content likely to be similar, and we understand services would already utilise such information for making recommendations where positive feedback is given.

20.196 **Testing and roll-out.** We also expect that there would be costs in doing a potentially significant amount of testing of this feature. Before rolling out this measure more widely, services are likely to need to run a variety of front-end and back-end tests. Service providers may also need to do a phased roll-out and gather and assess data as to its operation and effectiveness. We understand that there are different ways in which service providers could implement this functionality, and we are not being prescriptive in this measure to enable service providers who best understand their individual services and infrastructure to innovate in how this measure is designed.

20.197 Across these activities, we estimate that implementing the feature that allows the user to privately express negative sentiment and training the recommender system to process negative signals could take approximately 16 – 40 weeks of labour time split across roles including software engineers, machine learning engineers, applied scientists, and frontend designers and developers. This will need to be combined with an equal amount of non-software engineering time (e.g. project management, legal, trust and safety). Using our assumptions on labour costs required for this type of work set out in Annex 12, we estimate that the one-off direct costs could be somewhere in the region of £36,000 to £178,000. Although our cost estimates across all measures are based on the same salary assumptions, we recognise that this measure is complex and might require a particularly high level of expertise to implement. Therefore, salaries and hence costs for this measure may be less likely to be around the lower bound compared to some other measures.

20.198 The cost of this measure for a given service will be impacted by the existing design of the recommender system. We understand that there are several factors that are likely to cause costs to be higher for some services to implement the measure, but we have some uncertainty about exactly what features of service would cause costs to be towards the lower end of our estimate. Factors likely to affect implementation costs include increased system complexity (such as the number of models which feed into a recommender system, or different design architectures) and the number of users or languages in the system, which increase the number of variables and amount of data needing to be processed. As noted, some service providers have already developed similar features of negative feedback tools, and they would likely incur lower costs to adapt this feature to fit this recommendation compared to services that have no negative feedback functionality.

20.199 We understand there are likely to be some minor additional data storage costs from increasing the complexity of the recommender system but believe these are negligible. A recommender system that can receive and respond to negative user feedback is likely to require marginally more storage space.

20.200 There may be additional oversight and coordination costs associated with changing products. Larger businesses can be more complex and may have larger teams or different teams which need to communicate to implement changes to their services. We would

expect the oversight and coordination costs to be largely correlated with the size of the company, but do not have sufficient information to be able to quantify these.

20.201 In addition to the implementation costs, we would expect a service to incur ongoing costs including maintenance costs to ensure that this feature continues to function as intended. In line with our standard cost assumptions set out in Annex 12, we assume this to be approximately 25% of the initial set-up costs, ranging from approximately £9,000 to £44,000 per year. To maintain optimal performance, services fine-tune and test their recommender systems frequently. In cases where a service does not already have a negative feedback feature, when introducing negative user feedback, services may observe additional technical metrics on whether the system responds and adapts to negative user feedback on content. Making sure that the measure is working as intended in terms of how the function is being used, and how subsequent recommendations are impacted by negative feedback is likely to be a more involved process. Consequently, we anticipate that this measure may increase the time to run and review existing testing procedures. Therefore, for this measure we consider that the maintenance costs may be towards the upper end of our estimates, as unlike other measures there may be more labour time needed to continue to ensure the feature remains effective over time and is able respond to new types of content being recommended and new users using the tool.

Costs to services resulting from age assurance

20.202 Services in scope of this measure may apply highly effective age assurance to target this measure at children, although this is not specifically recommended. The costs of highly effective age assurance are covered in the Age Assurance, Section 15 within this Volume.

20.203 Services in scope of this measure will necessarily also be in scope of either Measure RS1 or Measure RS2 in this section. Therefore, any costs related to age assurance would already be captured under those measures.

Indirect costs to services

20.204 We consider that it is likely that some non-harmful content will not be recommended to children due to this measure, and we have considered the potential for this to have some business impact where this content would have been engaging to users.[820] This would occur when a service limits the prominence of non-harmful content that a user would have found engaging. However, this is outweighed by the cases where this measure reduces the prominence of content that children prefer not to see, which may make them more engaged with a service due to the reduced risk of encountering content they do not want to see. In addition, we believe that the risk of reduced engagement is limited, as services can recommend a wide range of other non-harmful content instead.

[820] For more detail, please see Business models and commercial profiles set out in volume 3, sub-section 7.12.

Which providers we propose should implement this measure

20.205 We propose to recommend this measure for all large U2U services likely to be accessed by children that have a content recommender system (as defined in Box 1), and are medium or high risk for at least two kinds of PPC, PC (except bullying content), and the relevant kinds of NDC that we are minded to apply it to (body image and depressive content, subject to the outcome of the consultation). [821] [822]

20.206 Recommender systems are a key pathway for children to encounter content harmful to children, and yet children do not feel they have control over the content they are shown. This measure enables children to say when they want to see less of specific kinds of content being recommended to them. By using this information to limit the prominence of recommendations of similar content to that child, this measure is intended to help reduce the exposure of children to harmful content. We also believe that this measure may provide information to a service about content which could be NDC and can act as a safety net if there is still some residual harmful content in recommender feeds after Measures RS1 or RS2 have been implemented.

20.207 Given these material benefits to children, we consider it proportionate to recommend this measure for large services with medium or high risk for two or more types of relevant content. We consider that the measure has the potential to improve children's safety in relation to all kinds of content harmful to children which we have evidence are linked to recommender systems. These services are more likely to have both high numbers of children on the service and significant volumes of harmful content that could appear in children's recommender feeds, which Measures RS1 and/or RS2 may not adequately address by themselves.

20.208 We are also confident that large services will typically have access to the necessary technical skills to implement this measure effectively, and that the costs imposed by this measure will still be a relatively small increase in their costs. As set out in the 'Effectiveness' sub-section above, we understand that many large services have experience in developing sophisticated recommender systems capable of responding to real-time user feedback and adjusting future recommendations accordingly. This is likely to improve the effectiveness of the measure in correctly actioning a child's intent as to what they do not want to see more of.

20.209 However, we do not consider it proportionate to recommend this measure for services that are not multi-risk. For services with low risk of content harmful to children this measure is likely to have limited benefit, if any, to children's online safety.

---

[821] We expect services to take appropriate action against bullying content in relation to the proposed measures set out in Content moderation for U2U services and Measure RS3. However, given limited evidence of the connection between bullying content and recommender systems, a service would not be in scope of the measure if the children's risk assessment does not identify any risk of PC other than bullying on the service and there is no other risk of harm. If we receive more evidence of the connection between bullying and recommender systems, we may be able to revise this position. As explained in the draft Children's Register of Risks, Section 7.9, our preliminary assessment is that body image content and depressive content could meet the definition of NDC, subject to further evidence on defining these types of content and establishing the link to significant harm. If confirmed, these would be included within the risk criteria for this measure. In the future, should new types of NDC be identified, we will consider the connection to recommender systems before assessing whether these would also bring a service into scope of this measure.

[822] This is similar to the concept of 'multi-risk' but limited to risks of kinds of content that are linked to recommender systems, as explained in the previous footnote.

20.210 For services with medium or high risk for a single relevant kind of content, which would in any case be in scope of Measure RS1 or Measure RS2, we consider that those measures should already provide significant protections, whereas the incremental benefit from Measure RS3 would be greater where services are seeking to mitigate significant risks across several kinds of harmful content and can better deal with this challenge by making use of negative sentiment signals, in addition to Measures RS1 and/or RS2. Additionally, while the benefits of this measure in providing greater protection to children online scale with the number of risks posed by a service, the cost to implement would not vary.

20.211 We also do not consider it proportionate to recommend this measure for smaller services, even where they pose relevant risks. In reaching this view we have considered that there are relatively high costs associated with implementing this measure in addition to the costs of Measures RS1 and/or RS2 which such services would be recommended to apply. Costs are uncertain and it is possible that some smaller services could incur costs above the lower end of our estimated rate. Also, while technically feasible, we understand that a personalised negative feedback loop can be complicated to implement, and so we are not confident that smaller services would be able to implement it with a reasonable level of cost in proportion to the benefits it would bring. We have considered the combined implications for this measure on top of the other two measures in this section, as well as the wider package of proposed measures, as discussed in our Combined Impact Assessment (see Section 23 within this Volume). In doing so, we have prioritised Measures RS1 and RS2 for smaller services where we believe that the benefits are more material.

20.212 In addition, we have concerns that small services would find it harder to manage potential unintended consequences, which could undermine the benefits of this measure or lead to additional adverse impacts. For instance, this could occur if increased friction worsens the user experience, or if users stop using the feature if they do not feel that their negative sentiment is having the intended impact on their recommendations.

20.213 Overall, at this time we consider that Measures RS1 and/or RS2 are more likely to adequately protect children on smaller services, and Measure RS3 would be disproportionate for those services. We will continue to collect evidence on the effectiveness and costs of the measure at preventing children from encountering harmful content and may consider extending it in future iterations of the code.

20.214 Therefore, we have provisionally concluded this measure should apply to all large U2U services likely to be accessed by children that have a content recommender system, and are medium or high risk for at least two kinds of content that is PPC, PC except bullying content, or the relevant kinds of NDC that we are minded to apply it to (body image and depressive content), subject to the outcome of the consultation.

Provisional conclusion

20.215 Given the harms this measure seeks to mitigate in respect of content harmful to children as well as the risks of cumulative harm recommender systems pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Draft Children's Safety Codes. For the draft legal text for this measure, please see PCU F3 in Annex A7.

21. User support measures

Functionalities such as group messaging and commenting on content play an important role in the way that services operate and present numerous benefits to children including staying in touch with, and feeling connected to, family and friends. However, these functionalities can also allow other users to add children, without their consent, to group chats that distribute harmful content, and can expose children to harmful content in comments or by messaging them directly. Our Children's Risk Profiles identified these functionalities as posing risks to children. The measures proposed in this section are designed to address the risks from these functionalities.

We propose six user support measures in total.

Three of our six recommendations are for user support tools. These would help to protect children by giving them more control over their online experience.

The remaining three recommendations focus on making supportive information available to children. These would mitigate the impact of harmful content that children may encounter online and help children to understand the user tools available to them.

Our proposals for user support tools seek to prevent children from encountering harmful content such as pornographic content, suicide, self-harm or eating disorder content, bullying content, abuse and hate content and violent content.

Our proposed recommendations for supportive information would help to mitigate the impact of different kinds of harmful content by signposting children to supportive information and helping them to understand what action they can take if something goes wrong.

Our proposals

#	Proposed measure	Who should implement this[823]
User support tools		
US1	Provide children with an **option to accept or decline an invite to a group chat**	All U2U services that: • Have group chats, and • Are medium or high risk for one or more of: pornographic content, eating disorder content, bullying content, abuse and hate content or violent content
US2	Provide children with the **option to block and mute other users' accounts**	All U2U services that: • Have user profiles and certain user interaction functionalities, and • Are medium or high risk for one or more of: bullying content, abuse and hate content or violent content
US3	Provide children with the **option to disable comments on their own posts**	All U2U services that: • Have comment functionalities, and • Are medium or high risk for one or more of: bullying content, abuse and hate content or violent content

[823] These proposed measures relate to providers of services likely to be accessed by children.

#	Proposed measure	Who should implement this[823]
Supportive information		
US4	The provision of **information** to child users **when they restrict interactions with other accounts or content**	All large U2U services that: • Are multi-risk for content harmful to children
US5	**Signpost children to support** at key points in the user journey	**Intervention point 1 when children report content:** All U2U services that: • Are medium or high risk for one or more of: suicide content, self-harm content, eating disorder content or bullying content. **Intervention point 2 when children post or re-post content**: All large U2U services that: • Have posting/re-posting functionalities, and • Are medium or high risk for one or more of: suicide content, self-harm content, eating disorder content or bullying content, and • Have measures that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder or bullying content. **Intervention point 3 when children search for harmful content:** All U2U services that • Have user-generated content searching, and • Are medium or high risk for one or more of: suicide content, self-harm content or eating disorder content, and • Have measures that enable them to become aware of when a user searches using suicide, self-harm or eating disorder related search terms
US6	**Provide age-appropriate user support materials** for children	All U2U and search services that: • Are multi-risk for content harmful to children

Consultation questions

53. Do you agree with the proposed user support measures to be included in the Children's Safety Codes? Please confirm which proposed measure your views relate to and provide any arguments and supporting evidence. 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.

What are user support measures?

21.1 In this section, we outline two categories of user support measure:

- User support tools: Measures US1, US2 and US3 are designed to give children appropriate control over who they interact with and what they see online; and
- Supportive information: Measures US4, US5 and US6 are designed to ensure that if something goes wrong online, children understand the user tools available to them and can access appropriate support.

21.2 All of our proposed measures apply to certain U2U services. Measure US6 additionally applies to certain search services, while Measure US5 has an equivalent measure that we are proposing in the Search features, functionalities and user support Section 22.

21.3 The Online Safety Act 2023 ('the Act') requires U2U service providers, where proportionate, to take or use measures in a number of areas in order to fulfil their duties to keep children safer online.

21.4 These areas include functionalities allowing for user control over content that is encountered, especially by children.[824] Our proposed measures specifically address group messaging, blocking and muting, and commenting functionalities. Implementing these measures should help providers of U2U services meet the duties to manage and mitigate the risks and impact of harm to children on the service, and to prevent or protect children from encountering content that is harmful to them.

21.5 Another area is user support measures, which may include the provision of supportive information and materials. Taking such measures should help U2U service providers meet the duty to manage and mitigate the risks and impact of harm to children on their service.[825] [826]

Definition box 1: Glossary of key functionalities	
Measure US1	
User groups	Online spaces that are often devoted to sharing content surrounding a particular topic or bringing together a community or group with shared links and interests (e.g. family groups, groups of parents, etc). Some groups have more than a thousand members.[827] User groups are generally closed to the public and require an invitation or approval from existing members to gain access. In some cases, they may be open to the public.
Group messaging	A functionality allowing users to send and receive messages through a closed channel of communication to more than one recipient at a time.
Measure US2	

[824] Section 12(8)(f) of the Act. There is an equivalent duty for search services in section 29(4)(c) of the Act, which is addressed by the measures we propose in the Search features, functionalities and user support Section 22.
[825] Section 12(8)(g) of the Act. Our proposed Measure 6, and measures presented in the Search features, functionalities and user support Section 22 are also recommended for compliance with the equivalent duty for search services in section 29(4)(e) of the Act.
[826] Section 12(2) of the Act. Our proposed Measure 6 would also help search services to meet their equivalent duty to mitigate and manage the risks and impact of harm to children, as laid out in section 29(2) of the Act.
[827] whatsapp.com, How to create and invite into a group. [accessed 05 March 2024].

Blocking	A user tool that allows a user (User A) to limit the interaction with another user (User B) or content from that user, so that: • User B cannot send direct messages to User A and vice versa. • User A will not encounter any content posted by User B on the service (regardless of where on the service it is posted) and vice versa. That content could include (but is not limited to) reactions to and ratings of content by User B; and content originally posted by User B and then reposted by another user. • User A and User B, if they were connected, will no longer be connected.
Muting	A user tool to allow User A to manage content they are served from User B, so that: • User A will not be served any content posted by User B on the service (including reactions to and ratings of content posted by User B, and content originally posted by User B and then reposted by another user). User A can still view content posted by User B by visiting their user profile. • Like blocking, User A will not be served content from User B. However, unlike with blocking, User B can contact User A and vice versa. Also, User B can see User A in search results, visit User A's page and see their content, and User A can visit muted User B's page and see their content. • Comparing to blocking, User B is much less likely to discover that User A has muted them, as their ability to find or directly contact User A is not reduced. Muting another user in this context is different from 'chat/group muting', which turns off notifications on a particular chat or group.[828]
Measure US3	
Commenting	A functionality allowing users to reply to content, or post content in response to another piece of content, which is then visible alongside the original content. Commenting can be on content that is accessible to the public through a U2U service or content that has been distributed within closed user groups.
Measure US4	
Content restriction tools	User tools that allow users to privately (i.e., not visible to any other user of the service, including the creator of the content) restrict their interaction with a piece of content or kind of content, so that less or none of that content appears in their content feed in future. In some cases, the user may still be able to access the content if they search for it directly. These tools have different names on different services. Examples we are aware of include 'see less of this' and 'hide' tools. We would not consider a 'dislike' button to be a content restriction tool, if its primary function is to publicly express an opinion about the content, not to restrict interaction with it. However, a 'not interested' button might be a content restriction tool for the purposes of this measure if its primary function is to allow users to privately restrict interaction with a piece or kind of content.

[828] We recognise that many services do also allow muting on chats, however this not in scope of our measure.

Why are user support measures important for protecting children?

Functionalities

21.6 Our evidence suggests that some functionalities can present risks of children encountering content that is harmful to them.[829] Our analysis of the causes and impacts of harms to children in Section 7 details how different functionalities may contribute to the risk of children encountering different kinds of harmful content, as defined in the Act:

- For services that have user groups with group messaging functionalities, we have identified children as being at risk of pornographic content, eating disorder content, bullying content, abuse and hate content and violent content.
- On services that allow users to connect with one another we find children to be at risk of bullying content, abuse and hate content and violent content.
- Our evidence also demonstrates that commenting functionalities increase the likelihood of children encountering bullying content, abuse and hate content, and violent content.[830]

Information

21.7 Evidence suggests children are safer online when service providers give them choices and the information necessary to make those choices. Making information about a service and its functions clear and accessible to children is key to ensuring children can understand and digest its contents and take action when something goes wrong. Responses to our 2023 Protection of Children Call for Evidence (our 2023 CFE), as well as existing guidance, demonstrate that providing children with relevant, engaging and comprehensible information and support is vital in keeping them safe online.[831] [832] In particular, evidence suggests that providing children with supportive information at key points in the user journey can help mitigate the impact of harm caused by suicide, self-harm, eating disorder and bullying content.[833]

21.8 Children cannot benefit from the protections offered by user support tools and other safety features if they do not know that they exist, or how to use them. Providing materials designed for and targeted at children and the adults who care for them should avoid this problem, ensuring that children have the knowledge and confidence to employ these tools online, mitigating the risks and impacts of harm caused by content that is harmful to them.

21.9 Another category of user support measures is parental controls. We are not proposing parental controls in the consultation but explain in Section 13, Overview of the Codes, that this may be a future focus of our work.

---

[829] See Box 1: Glossary of key terms for definitions of the functionalities that we address in this section.
[830] Section 7.1, Pornographic content; Section 7.3, Eating disorder content; Section 7.5, Bullying content; Section 7.4, Abuse and hate content; Section 7.6, Violent content.
[831] Molly Rose Foundation, 2023. Molly Rose Foundation response to our 2023 Protection of Children Call for Evidence. Anti-Bullying Alliance, 2023. Anti-Bullying Alliance response to our 2023 Protection of Children Call for Evidence; Samaritans, 2023. Samaritans response to our 2023 Protection of Children Call for Evidence.
[832] See for example, ICO, 2020. Age appropriate design: a code of practice for online services [accessed 16 April 2024].; and Designing for Children's Rights, 2022. Design Principles: Version 2.0. [accessed 16 April 2024].
[833] See paragraphs 21.187 - 21.192 below.

Interaction with Illegal Harms

21.10 In our Illegal Harms Consultation we proposed the following user support measures in our draft code for U2U services:

21.11 **Measure 7B**: Supportive information is provided to children using a service in a timely and accessible manner at various points in the user journey. This is to help child users make informed choices about risk by giving them information which could include access to safeguarding processes and support on a service. We have proposed this measure for services with relevant functionalities that are at risk of grooming.

21.12 **Measure 9A**: Users are able to block or mute other specific users and all user accounts which they are not connected to. We have proposed this measure for large services with relevant functionalities that are assessed as medium or high risk for at least one of various Illegal harms.[834]

21.13 **Measure 9B**: Users can disable comments relating to their own posts, including comments from users that are not blocked. We have proposed this measure for large services with relevant functionalities that are assessed as medium or high risk for at least one of various Illegal harms.[835]

21.14 We provisionally consider that the above measures as proposed in the draft Illegal Content Codes are also proportionate for providers of services likely to be accessed by children in relation to their additional duties for the protection of children. The measures proposed in the draft Illegal Content Codes were designed to protect users from various Illegal harms and to protect children from grooming. We therefore propose that these measures be included in the draft Children's Safety Codes as well.

21.15 Our evidence shows that giving children the option to block or mute other users and to disable comments on their own posts can also protect them from content that is harmful to children, including bullying content, abuse and hate content and violent content. We therefore propose that these measures be included in the Children's Safety Codes as well.

21.16 In our Illegal Harms Consultation, we also proposed Measure7B: Support for child users. We have proposed this measure for all U2U services that identify as high risk of grooming, or large services that identify as medium risk of grooming. While this measure would offer children some protection, we are also proposing a measure for large U2U services that are high or medium risk of at least two kinds of content harmful to children that supportive information should be provided when a child user takes action to restrict content and limit interaction with users (Measure US4).

Our proposals to protect children

21.17 The measures that we are proposing in this section fall into two categories:

21.18 **User support tools**: Measures US1, US2 and US3 are designed to give children appropriate control over who they interact with and what they see online; and

---

[834] Grooming; encouraging or assisting suicide (or attempted suicide) or serious self-harm; hate; harassment, stalking, threats and abuse; controlling or coercive behaviour.

[835] Grooming; encouraging or assisting suicide (or attempted suicide) or serious self-harm; hate; harassment, stalking, threats and abuse.

21.19 **Supportive information**: Measures US4, US5 and US6 are designed to ensure that if something goes wrong online, children understand the user tools available to them and can access appropriate support.

21.20 To meet the children's safety duties, the Act requires U2U services likely to be accessed by children to take or use measures in a number of areas, where proportionate. These include:

a) Functionalities allowing for control over content that is encountered, especially by children.[836] Our proposed user support tools relate to this area.

b) User support measures.[837] Our proposed supportive information measures relate to this area.

Summary of proposed measures and which services they apply to

21.21 In this section we discuss the following proposals:

- **Measure US1:** We recommend all U2U services that have group chats **and** are medium or high risk for one or more of pornographic content, eating disorder content, bullying content, abuse and hate content and violent content provide children with an option to accept or decline an invite before being added to group chats.
- **Measure US2:** We recommend all U2U services that have user profiles and certain user interaction functionalities **and** are medium or high risk for one or more of bullying content, abuse and hate content and violent content provide children with the option to block or mute other user accounts on the service.
- **Measure US3:** We recommend all U2U services that have comment functionalities **and** are medium or high risk for one or more of bullying content, abuse and hate content and violent content provide children with the option of disabling comments on their own posts.
- **Measure US4:** We recommend large U2U services that are multi-risk for content harmful to children provide children with supportive information when they take action against another user or kind of content.[838] This should include information about the effect of the action and other actions they may take to protect themselves further. This could include information on reporting, blocking or muting tools, among others.
- **Measure US5:** We recommend all U2U services that are medium or high risk for one or more of suicide, self-harm or eating disorder content, or bullying content signpost children to appropriate support when they encounter that content at key points in the user journey.[839]
- **Measure US6:** We recommend that U2U and search services that are multi-risk for content harmful to children provide age-appropriate user support materials, clearly explaining to children the user tools available to them on the service.

[836] Section 12(8)(f) of the Act. The equivalent duty for search services can be found in section 29(4)(c) and is addressed in our codes in the Search features, functionalities and user support Section 22.
[837] Section 12(8)(g) of the Act. The equivalent duty for search services can be found in section 29(4)(e) and is addressed in Measure 6 in this section and in the Search features, functionalities and user support Section 22.
[838] See Framework for Codes at Section 14 within this Volume for a definition of a large service.
[839] See details on the measure for the relevant functionalities and intervention points relevant to different services.

Which users these measures apply to

21.22 Where U2U services implementing Measures US1 to US5 are using highly effective age assurance (including where they are implementing the measures we propose in Age Assurance Section 15), they can apply these measures only to child users, although they can choose to apply them to all users if they wish.

21.23 Where these services are not using highly effective assurance, they should apply Measures US1 to US5 to all users.

21.24 If a service uses highly effective age assurance to apply these measures only to child users, it may incur additional cost. However, since we are not recommending this, and it is optional, we do not consider these costs in our assessments of the implications for services for each of Measures US1 to US5.

21.25 U2U and search services that are multi-risk for content harmful to children should apply Measure US6 to all users. These materials should be available to users and non-users of a service so do not require the use of highly affective age assurance to apply them to children only.

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

Explanation of the measure

21.26 This measure is designed to prevent children being added to group chats by others when they do not want to be. It recommends that services which offer group messaging functionality provide children with a message prompting them to either accept or decline an invite to join a group. In this discussion we frequently refer to user groups in services which offer group messaging functionality as 'group chats' for simplicity.

21.27 When another user attempts to add a child to such a group, including a group that the child has previously declined to join or chosen to leave, the child should not be added immediately and the service should send the child a message, for example a notification or prompt, informing them of the request to add them.

21.28 The message should include any relevant publicly visible information about the user inviting the child to join, as well as such information about the group. A potential example would be the group's name, the account name of the user who invited the child, any description of what the group is, the date it began, and the number of members. This would help to inform the child's decision about whether to join the group chat. Until the child has responded by accepting or declining, or the message has expired, it should remain easy for the child to find.

21.29 The message should be clear, use language that is comprehensible for children, and should be neutral. It should not nudge the child to accept a request, for example by making the accept option easier or more prominent. Evidence suggests that the design of safety measures can influence their effectiveness, for example in relation to the impact of default

settings on the effectiveness of alerts among adults, and we would expect that this also applies to children.[840]

21.30 Services should allow the child a reasonable time to respond to the message. It is up to services to decide whether the message should expire automatically after a reasonable period if the user does not respond or should not be time limited at all.

Effectiveness at addressing risks to children

21.31 Services that offer one-to-one connections typically provide users with an option to accept or decline an invitation to connect. However, on some services it is possible for children to be added to group chats without being given this option. For example:

- **Discord:** Users can add friends to group chats directly.[841]
- **Instagram:** Teens can only be messaged or added to group chats by people they already follow or are connected to.[842] Teens in supervised accounts need permission from a parent to change this setting.[843] This default setting will apply to all under-16s (or under 18s in certain countries).
- **Snapchat:** Only friends can add other users to group chats.[844]
- **WhatsApp:** Users can choose between three options: allowing everyone, only contacts, or specific contacts to add them to group chats. The default setting is 'everybody'.[845]

21.32 It may also be possible for users to be re-added without their agreement to a group chat they have left.

21.33 Evidence in our draft Children's Register of Risk shows that various kinds of harmful content are shared in groups, meaning that group messaging facilitates children's exposure to PC and PPC (Governance, Systems and Processes, Section 7.11). This can be exacerbated when children are added to a group by someone else. In relation to bullying content, children reported that they could be targeted in group chats to which they had been added without giving permission. Children also reported being added to groups with users they had previously blocked, though research participants said some services would notify them to check they were aware before doing so.[846]

21.34 Research suggests that being added to group chats by others can lead to children encountering violent and abusive content, with participants suggesting that services should give users the option to accept a group chat invite.[847] Being added to group chats by others

---

[840] Ofcom, 2023. Default effects and alert messages The impact of auto-play and auto-skip defaults on the effectiveness of alert messages.
[841] Discord.com., Group Chat and Calls. [accessed 29 February 2024].
[842] Meta, 25 January 2024. Introducing Stricter Message Settings for Teens on Instagram and Facebook. [accessed 17 April 2024].
[843] Supervision on Instagram is a set of tools and insights that parents and carers can use to help support their teens (ages 13-17). Supervision is optional, and both the parent or carer and the teen must agree to participate. It can be removed at any time by either person. The other person will be notified that supervision has been removed. See Instagram.com., About supervision on Instagram. [accessed 17 April 2024]
[844] snapchat.com., How do I add Snapchat friends to a Group Chat?. [accessed 29 February 2024].
[845] whatsapp.com., How to change group privacy settings. [accessed 29 February 2024].
[846] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.
[847] Ofcom 2022. Research into risk factors that may lead children to harm online.

could also lead to children encountering pornographic and illegal content.[848] Children have witnessed young people sharing sexual content on a group chat, profile or forum (Governance, Systems and Processes, Section 7.11). In some cases, this may amount to an illegal harm such as cyberflashing or where pornography is shared with a child as part of the grooming process including to persuade a child to share self-generated indecent imagery (SGII).[849] [850] [851] Some children may also witness people sharing child sexual abuse material, including SGII, on group chats. The proposed measure may also help to prevent children from encountering such illegal content via such group chats. It also complements the measures for default settings for children that we have proposed in our Illegal Harms Consultation, which add friction to communication routes for perpetrators to target children. The proposed measures would further restrict perpetrators' ability to contact children by other means.[852] The benefits of the proposed measure in protecting children from harmful content are sufficient on their own, but we consider that these extra benefits relating to illegal content are valuable as well.

21.35 Among children who have encountered content they felt promoted, glamourised or romanticised eating disorders, this is often shared through messaging services. Children and young people with lived experience reported that sharing of content of this nature typically occurs in closed groups on messaging and social media services.[853] A study from 5Rights illustrated how children engaging with weight-loss content could then be added to messaging groups where extreme disordered eating behaviours were encouraged.[854]

21.36 We do not have sufficient evidence to suggest that children may encounter self-harm content as a result of being added to group chats by others, to recommend that services who assess as being at risk of self-harm content adopt this measure.

21.37 However, where services adopt it, the proposed measure may nevertheless help to prevent children encountering PPC. This is because the proposed measure would apply to all group chats, and not only to group chats where specific kinds of harmful content might be shared. Where others try to add them to any group chat, it may therefore give children some indication of the kind of content and users they are likely to encounter and decide whether

---

[848] 5Rights Foundation, 2021, Pathways: How digital design puts children at risk. [accessed 23 June 2023]. Note: The research involved setting up a series of avatars, which were profiles set up on social media apps that mimicked the online profiles of real children who took part in the interviews for this project. The age of the real child was used to register the profile and displayed in the bio of the user account.

[849] Childnet, 2017, Young People's Experience of Online Sexual Harassment. [accessed 6 February 2024].

[850] Section 6M, Intimate Image Abuse, of our 2023 Illegal Harms Consultation.

[851] Section 6C, Child Sexual Exploitation and Abuse (CSEA), of our 2023 Illegal Harms Consultation.

[852] Section 18, U2U default settings and support for child users, of our 2023 Illegal Harms Consultation.

[853] In this study, we spoke to children and young people who have encountered content they felt promoted glamourised, or romanticised eating disorders, self-harm, or suicide. We also spoke to children and young people who had encountered the content and who also had lived experience of an eating disorder, self-harm, suicidal ideation, anxiety, and depression. Ofcom, 2024. Online Content: Qualitative Research Experiences of children encountering online content relating to eating disorders, self-harm and suicide.

[854] The study, which involved interviews with children, describes the experience of a child who was concerned about their weight and started searching for weight-loss tips and diets on social media. After following 'thinspiration' accounts and posting about her weight loss, she soon connected with a community of users engaging with similar content and was added to several messaging groups. These groups encouraged extreme dieting and users requested verbal abuse to hold them to account on their disordered eating behaviours. 5Rights Foundation, 2021. Pathways: how digital design puts children at risk. [accessed 2 August 2023]

they join. There is awareness among some children of the risks posed by group chats, particularly those that include people they do not know.[855]

21.38 On services that do not already offer the option to choose before joining a group chat, this proposed measure would create an opt-in step that introduces user choice and control. Research suggests that children value having control over their online experiences, including who can contact them.[856] It would help to protect children from encountering at least pornographic content, bullying content, abusive content, content which incites hatred and violent content, as well as helping to prevent them encountering eating disorder content.

21.39 We consider this proposed measure is technically feasible. It is similar to existing practices to give users an option to accept or decline an invitation to connect individually from another user, such as a friend request, as well as the options that services currently offer, described above, that give users control over who can add them to a group chat.

Rights assessment

21.40 As set out above, evidence shows that group functionalities bring about children's exposure to harmful content and the consequences of such exposure can include significant harm to children's physical, mental or emotional wellbeing.[857] We expect this proposed measure would help ensure that children are protected from encountering content that is harmful to them, in line with the legitimate objectives of the Act.

21.41 In implementing this measure there is however a potential impact on the rights of users (including of children and adults), in particular, their rights to privacy (Article 8 of the ECHR), freedom of expression (Article 10 of the ECHR) and freedom of association (Article 11 of the ECHR).[858] We have considered the extent to which the degree of interference with these rights is proportionate. We recognise that the Act requires services to use proportionate systems and processes to prevent and protect children from encountering content that is harmful to them. We therefore consider that a substantial public interest exists in measures which aim to prevent and protect children from encountering harmful content, in the protection of children's health and morals, public safety and in particular the protection of the rights of others, namely child users of regulated services.

21.42 Where services implementing this measure use highly effective age assurance, they can apply this measure to only children or they should apply them to all users (i.e., both children and adults). Where services implementing this measure are not using highly effective age assurance, then both adult and child users would be subject to the measures. We therefore

---

[855] London: National Crime Agency and Brook, McGeeney, E. & Hanson, E. (2017). Digital Romance: A research project exploring young people's use of technology in their romantic relationships and love lives. [Accessed 17 November 2023].

[856] NSPCC, 2017. Net Aware Report 2017: "Freedom to express myself safely". A summary of the results of a large-scale study examining the opportunities and risks experienced by young people in their online lives. [Accessed 17 November 2023].

[857] See the draft Children's Register of Risk, Section 7.

[858] We note that adult users may include those who are operating on behalf of a business, or accounts that might also be concerned with other entities, such as charities, as well as those with their own, individual account. Both corporate and individual users can benefit from the right to freedom of expression, and we acknowledge the potential risk of interference with the rights of these users to freedom of expression.

have considered the impact of such measures on both child users and adult users' rights to freedom of expression and freedom of association and privacy.

Freedom of expression and association

21.43 As explained in Volume 1, Section 2, Article 10 of the ECHR upholds the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 11 of the ECHR upholds the right to associate with others. The right to freedom of expression and freedom of association are qualified rights. Ofcom must exercise its duties under the Act in light of users' and services' Article 10 and 11 rights and not restrict these rights unless it is satisfied that is necessary and proportionate to do so.

21.44 We acknowledge that this proposed measure may impact the online experience of child users, and potentially adult users' too if this functionality is extended to all users as noted above. This is because it involves services adding an extra layer to the user's journey when they are added to a group chat, which may create some friction and delay to the process by which users are normally added into a group chat. In addition, we acknowledge that this measure has the potential of adding a friction on how users (including both children and adults) express or share information and ideas where for instance a child user (or potentially an adult user if this functionality is extended to all users) refuses to be added to a group chat as a result of this measure.

21.45 In this way it could have some impact on children's (and potentially adults') ability to access content (including content that is harmful, but also non-harmful content which may include the most highly protected forms of speech, such as political expression). It may also therefore affect other users' ability to share content with children or to associate with others (to the extent that an invitation to join the group is declined). However, we consider that any such impact on rights of freedom of expression or association would be limited since the decision to refuse to join a group chat would be a choice made solely by the user concerned, and would have no impact on the right of other users to express their ideas and share information on the service concerned with any other users who are members of the group chat, or elsewhere on the service. In addition, we consider the risk that child users will not join group chats they would want to be in as a result of the potential increased friction, such as by missing a notification, is minimal as we are recommending the measure is clear, and users are given a reasonable time to respond, but are otherwise giving services flexibility to do so in the way that best meets the nature of their service and their users' needs.

21.46 We also consider that giving child users (or potentially all users) this functionality to increase their ability to make informed choices about what group chats they join could also have positive impacts on their freedom of expression and freedom of association rights, for example, children may feel more able to join online communities where they feel safe and receive and impart (non-harmful) ideas and information with other users in those groups.

21.47 Therefore, to the limited extent that this measure restricts children's (and potentially adults') ability to access and share content and associate with other users, we consider that this is justified and proportionate in line with the duties of the Act, as the benefits of the protections on children should outweigh the limited restriction on other users' rights to share content with children. While there might also be the potential for a minimal impact on services' rights to freedom of expression, as this measure would increase frictions in the way that users connect on the service, we also consider this is justified and proportionate for the reasons set out above.

Privacy

21.48 As explained in Volume 1, Section 2, Article 8 of the ECHR confers the right to respect for individuals' private and family life. An interference with the right to privacy must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest. Again, in order to be 'necessary', the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.

21.49 We do not expect this proposed measure to result in any interference with any user's rights to privacy under Article 8 ECHR; indeed, giving children the option not to join a group chat could have positive benefits for their privacy. We acknowledge that implementing the functionality to give users the option to join or decline an invitation to a group chat would likely involve the processing of users' personal data – for example, it may be necessary to identify the account who is extending the invitation to join a group so that the child user in question can make an informed decision about whether or not to accept it. However, we consider it likely that the amount of additional personal data processed by the service to implement this functionality, above and beyond what they would already be processing for the purposes of enabling users to be added to group chats, to be minimal, particularly given we are suggesting that only publicly available information should be presented. In addition, services would need to comply with data protection legislation in connection with any such processing of personal data. We believe that these measures are compatible with the ICO Children's code, as they are an age-appropriate application of common U2U service features and in that context, avoid using children's personal data in a way that might be detrimental to their safety and wellbeing.

21.50 We recognise the possibility that where the measure is only applied to children, a user's inability to add them automatically could in itself send a signal that the user may be a child. However, we took into account the other protections we are recommending for children, in particular regarding direct messaging, which should mitigate the potential for harm. Services should also ensure that they take steps to mitigate this potential concern, implementing it in a way that is compliant with data protection law and in particular, the ICO Children's Code. As such, we consider that this measure takes a proportionate approach to users' privacy rights.

21.51 We acknowledge that if services use highly effective age assurance to target this proposed measure at child users only, there could be privacy impacts associated with the use of highly effective age assurance, as discussed further in Age Assurance Section 15 of this Volume. However, as we would only anticipate that services would be likely to apply highly effective age assurance for the purposes of this measure where they are already deploying it for other reasons (for example, to implement the measures set out in Age Assurance Section 15 of this Volume), we do not consider there would be any additional privacy impacts as a result of use of high effective age assurance in connection with this proposed measure.

Impacts on services

21.52 Below we discuss the direct costs to services from implementing the measure and potential indirect costs.

Direct costs of implementation

21.53 We would expect services that are not currently providing children with the option to accept or decline group invites to incur direct costs of software engineering time to modify their service in line with this proposed measure. We expect that costs would be significantly lower

where a service already provides an optional feature for users to accept or decline before being added to groups by others, and so have estimated costs in both cases as summarised in **Table 21.1.**

Table 21.1: Summary of direct cost estimates

Activity	One-off implementation cost	Ongoing annual cost
Implementing from scratch	£20,000 – £67,000	£5,000 – £17,000
Modifying existing functionality	£4,000 – £9,000	£1,000 – £2,200

Source: Ofcom analysis

21.54 Where a service does not have this type of functionality, we consider that the likely activities the provider would need to undertake to build this are: designing the feature in a way that is user-friendly, aligns with the overall design language of the platform and fits with existing features; building and testing the front-end and back-end features; and releasing the feature into the live environment and tracking progress. Depending on the current system architecture and user interface, changes could involve input from user experience or user interface designers, graphic designers, web designers, content teams, developers, and software engineers.

21.55 We estimate that implementing this could take approximately 9-15 weeks of time across a range of technical professionals, and we assume that there is an approximately equal amount of time from other professions (e.g. project management). Using our assumptions on labour costs set out in Annex 12, we estimate that the one-off direct costs could be in the region of £20,000 to £67,000.

21.56 The approach and associated costs may depend on whether a provider is using an off-the-shelf tool to build and maintain its service, where there may be ready-made features or plug-ins available, or if the service needs to modify the underlying code and site infrastructure. The cost would also depend on the current structure of their systems. For example, the incorporation of a new functionality is likely to be significantly cheaper if the existing systems already incorporate some level of privacy design (e.g. if privacy options to allow users control over who they interact with on the service are already set-up within existing backend databases) or are designed on a modular basis which separates individual functions into independent programming modules. We understand that costs for implementing this measure would typically increase with the size and complexity of the service, where larger services might require more robust solutions and extensive testing, requiring more development time. If the feature deals with large amounts of data (e.g. user profiles, activity logs), the technical solution might need to be more complex to handle the load efficiently. Also, a large, complex service might have more existing functionalities and systems to integrate with.

21.57 Where services have already built this functionality as an optional setting, services can implement the measure by ensuring that the function is on for children and cannot be turned off by children. For such services, we estimate that this could be closer to two weeks of software engineer time for the testing and set-up process. Even though the underlying functionality already exists, developers may need to test the feature thoroughly to ensure it works as expected for users and does not disrupt existing functionality. For example, the existing functionality might have settings or configurations specific to how it works for users

who currently have it enabled, or the functionality might interact with other parts of the service. Once testing and adjustments are complete, the change needs to be deployed and monitored. With an equivalent amount of time from other professions to the two weeks of software engineering time this leads to an approximate cost estimate of £4,000 to £9,000.

21.58 In addition to the one-off costs associated with this measure, we expect there to be ongoing costs to review and monitor it. If ongoing costs consist of approximately 25% of the original one-off cost on an annual basis, this would be approximately £5,000-£17,000 per annum for services building the feature from scratch, and £1,000-£2,250 for services that have this existing feature but need to ensure it is applied to children.

Potential indirect costs

21.59 The measure may add friction to interactions between users, potentially making group chat functionalities less attractive to some users and reducing usage of the service.[859] Depending on the business model of the service, a reduction in usage could lead to a reduction in engagement with the overall service, which could in turn reduce revenue to services.[860] However, some users may also increase engagement as they feel safer knowing they can control which groups they are added to.

21.60 To the extent that children do not join groups that they do not want to be in, and this impacts a service's revenue, we consider this to be the aim of the measure and entirely justified. We consider that there is only a minimal risk of reduced engagement from children not joining a group that they would have wanted to be in (for instance because, of missing the notification) because we have designed the measure to mitigate this risk.

Which providers we propose should implement this measure

21.61 Our evidence in the 'Effectiveness' sub-section sets out the pathway to harm, whereby children are added to groups by others, causing exposure to certain kinds of harmful content. This measure is designed to reduce the risk of children being added to group chats unwillingly. By providing a choice to children, this measure enables children to decline invitations to group chats which they believe may be harmful to them. Given that the measure targets a specific risk factor (group chat invites), we believe it can deliver distinct incremental benefits over and above other measures we propose in our Codes, which are more general in nature, or which target different specific risk factors.

21.62 This measure is recommended for services that offer group chats and meet specific risk criteria. Our evidence suggests that the main risks of being unwillingly added to group chats by others are related to pornographic content, eating disorder content, bullying content, abuse and hate content and violent content. To protect children from these risks, we are proposing to apply this measure where there is a medium or high risk for at least one of these kinds of content.

21.63 We also considered whether to include services with medium or high risk in relation to any further specific kinds of content, in particular suicide and self-harm content. Although we

---

[859] See, for example Gencheva, A., 2018. Consumer Perceptions to Friction in the Context of the Privacy vs Convenience Trade-Off – The Case of an Open Banking Consent Journey. *iSCHANNEL* 13 (1). [accessed 26 March 2024]

[860] Section 7.12, Business models & commercial profiles, sets out the relationship between engagement and revenue for U2U services.

have some evidence of children encountering suicide and self-harm content in user groups, because we have limited evidence of this occurring due to being added to groups without an option to accept or decline, we are not proposing at this time to extend the measure to services with risks of these kinds of content. However, as noted in the 'Effectiveness' sub-section, the proposed measure may also offer some protection against children encountering other kinds of content, as well as illegal content, if these exist on the services implementing the measure.

21.64 We have considered whether this measure is proportionate for services of all sizes that meet the relevant functionality and risks criteria and have provisionally decided that it is. As noted above, by targeting a specific risk factor, the measure can materially improve children's safety incrementally to other measures.

21.65 The estimated costs of this measure for service providers are such that we recognise the possibility that some small businesses implementing the measure could struggle to carry this cost, in combination with the cost of other measures they may be implementing. Services may be discouraged from offering group chats, and where group chats are integral to business models, it could discourage some services from serving UK users. This could harm users who benefit from accessing these services and using group chats. However, we consider that costs would typically scale with the size and complexity of the service, and so believe that smaller services are likely to incur costs towards the lower end of our range of direct costs estimates. On balance, we believe the measure is proportionate relative to the capacity of services which have group messaging functionalities.

21.66 Overall, while noting the potential for some adverse effects, we have provisionally concluded to recommend this measure to all U2U services likely to be accessed by children that have a group chat functionality and are medium or high risk for one or more of the following kinds of content that is harmful to children: pornographic content, eating disorder content, bullying content, abuse and hate content, and violent content.

Provisional conclusion

21.67 Given the harms this measure seeks to mitigate in respect of pornographic content, eating disorder content, bullying content, abuse and hate content and violent content we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU G4 in Annex A7.

Measure US2: Provide children with the option to block and mute other users' accounts

Explanation of the measure

21.68 Evidence suggests that services which have functionalities such as user connections, posting content, and user communication can present a risk of bullying content, abuse and hate content and violent content. Tools that enable children to block or mute other users can help to protect them from encountering such harmful content that other users may be posting or sharing. This is because when you block a user you cannot see their posts, and they cannot message you or interact with your posts, and when you mute a user you are not shown their posts.

21.69 We propose to recommend the following measures:

- **Measure US2a:** Individual account blocking and muting: U2U services should provide children with the ability to block and mute other user accounts individually.
- **Measure US2b:** Global blocking of any non-connected account: U2U services should provide children with a public user profile a clear and accessible means of making themselves uncontactable to any user they do not have a mutually validated connection (that is, a connection that both users have agreed to).

21.70 These measures can help to protect children from encountering bullying content, abuse and hate content and violent content perpetrated either by a specific person, or by other users in general, where they think they are at risk of being targeted because of one of the characteristics specified in the Act (race, religion, sex, sexual orientation, disability, and gender reassignment).[861]

21.71 Muting is a less drastic option than blocking, which would allow children to reduce the risk of encountering harmful content without the person they are muting being able to discover that such an action has been taken.

21.72 As discussed in 'Interaction with Illegal Harms' above, this measure mirrors Measure 9A ('Measure to give all users the ability to block and mute other user accounts') in our Illegal Harms Consultation. Some large service providers implementing this measure would also be implementing the equivalent Illegal Harms measure.

Effectiveness at addressing risks to children

21.73 Evidence suggests that children may encounter bullying content, abuse and hate content and violent content through interacting with other users on these services (Governance, Systems and Processes, Section 7.11).

21.74 This measure can play an important part in reducing the likelihood that children encounter harmful content from specific users, or from any user that they do not have a connection with and would help to mitigate the impact of harm to children.

21.75 The evidence suggests blocking and muting tools are an important and effective means of managing children's online experience and reducing the risk on U2U services. Ofcom cyberbullying research highlights some children's use of blocking as a temporary measure to remove themselves from a situation to avoid escalation. For example, the study shows that some children find muting a particularly useful mitigation for bullying as users were unlikely to know they were muted and so this was less likely to result in escalation.[862]

- Services including X, Facebook, Instagram, TikTok, LinkedIn, Snapchat, YouTube, Medium and Tumblr offer user blocking and/or muting tools.[863]
- Meta's app, **Threads**, allows users to "unfollow, block, restrict or report a profile on Threads. Any accounts blocked on Instagram will automatically be blocked on Threads."[864] Digital marketplaces such as **eBay**[865] also provide user blocking tools.

[861] Section 62(2) and 62(3) of the Act.
[862] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.
[863] Pen America, Online harassment Field Manual; Blocking, Muting and Restricting. [accessed 09 April 2024].
[864] Meta, 2023. Introducing Threads: A New Way to Share With Text. [accessed 09 April 2024].
[865] eBay. Blocking a buyer on eBay. [accessed 09 April 2024].

- Some services currently provide users with a functionality allowing them to globally block all non-connected accounts:
- **X** gives users the option to set replies only to those accounts that the user follows.[866]
- **Instagram** allows pre-emptive blocking of new accounts set up by the user of a blocked account, so blocked users cannot re-establish contact by making new accounts.[867] It also enables users to block all direct messages.
- **Facebook** limits messaging to friends and other mutually validated connections, including Facebook Dating and Marketplace. Users outside of these categories cannot send a message; they can only send a request to message.[868]
- **Discord** allows users to block direct messages from users that are not on their friends list.[869]

21.76 The effectiveness of blocking individual accounts may be limited in circumstances where the blocked user creates new accounts through which to continue targeting the user who has blocked them. As we noted in our Illegal Harms Consultation, this pattern of malicious and repeated targeting of users through the creation of multiple accounts is commonly used by perpetrators as part of harms such as coercive control and stalking. [870] [871] Perpetrators of bullying content, and abuse and hate content may also adopt such patterns. As with the equivalent measure in our Illegal Harms Consultation, this Measure US2b is intended to go some way to address and deter users who may persist and set up multiple accounts to bully or abuse children, for example.

21.77 This proposed measure complements Measure 7A (Non-connected accounts do not have the ability to send direct messages to children using a service), which we have proposed in our Illegal Harms Consultation. This proposed measure would mean that in addition to being able to disable direct messages, children would be able to decide not to see content from those they have not made a connection with.

Rights assessment

21.78 As set out above, evidence shows that user connections facilitate children's exposure to harmful content and the consequences of such exposure can include significant harm to children's physical, mental or emotional wellbeing.[872] We expect that this measure would result in a reduction in the likelihood of children encountering content that is harmful to them and would mitigate the impact of harm to children presented by harmful content present on the service which is one of the core objectives of the Act. As with Measure US1, we consider below the potential impacts on users' rights to freedom of expression and association, and privacy. As with Measure US1, services may apply this measure to child users only where they use highly effective age assurance to identify child users, or else would need to apply this measure to all users (i.e. including adult users), and we have therefore assessed the potential impact under both scenarios.

---

[866] Twitter, 2020. New conversation settings, coming to a Tweet near you. [accessed 09 April 2024].
[867] Meta, 2021. Introducing new tools to protect our community from abuse. [accessed 09 April 2024].
[868] Meta, Control who can send messages to your Messenger Chats list. [accessed 09 April 2024].
[869] Discord, 2022. Blocking & Privacy Settings [accessed 09 April 2024].
[870] Paragraph 20.25, Illegal Harms Consultation.
[871] Refuge, 2021. Unsocial Spaces. [accessed 30 August 2023].
[872] See the Draft Children's Register of Risk, Section 7

21.79 We acknowledge that this proposed measure could limit the extent to which some users may be able to share their content with any user who blocks or mutes them, including by deploying global blocking.

Freedom of expression and association

21.80 While the rights to freedom of expression and freedom of association include the right to receive and impart information, and to associate with others, they do not include a right to compel others to listen to or to associate with you when they do not wish to. Affected users would not be prevented from receiving or imparting information or ideas by means of the service beyond the user that has chosen to block or mute them. We therefore consider that, to the extent that this proposed measure interferes with users' rights to freedom of expression or association, any such restriction is limited. We also consider that this measure could have positive benefits for children's rights to freedom of expression and association, as giving them the option to block or mute users with whom they don't wish to connect, may make them feel more able to receive and impart (non-harmful) ideas and information with other users on the service or join online communities where they feel safe.

21.81 Therefore, to the limited extent that this measure restricts children's (and potentially adults') ability to access and share content and associate with other users, we consider that this is justified and proportionate in line with the duties of the Act. While there might also be the potential for a minimal impact on services' rights to freedom of expression, as this measure would increase frictions in the way that users connect on the service, we also consider this is justified and proportionate for the reasons set out above.

Privacy

21.82 As with Measure US1 above, we do not consider this measure would interfere with users' rights to privacy, and indeed might have positive benefits for children's rights to privacy in that it would give them additional options for deciding how to share their personal information and content online. We also have identified similar data protection impacts as for Measure US1 above, as we expect that to the extent it is necessary for providers to process users' personal data to give effect to this, they must do so in compliance with data protection requirements.

21.83 We also consider that the impacts on user's rights to privacy with regards to the use of highly effective age assurance to target this measure at child users only would be similar to those in relation to Measure US1.

Impacts on services

21.84 Below we discuss the direct costs to services from implementing the measure and potential indirect costs.

Direct costs of implementation

21.85 For both Measure US2a and US2b, relevant services that are not currently implementing these measures would incur direct one-off costs to make the system changes to enable muting and blocking functions, and there would also be ongoing costs of maintaining these changes. The detailed assumptions underlying our direct cost estimates are found in Annex 12.

21.86 We estimated in our Illegal Harms Consultation that this type of functionality for both options is likely to require a one-off cost in the region of 20 to 150 days of software engineering time, with potentially up to the same again in non-engineering time. Making

assumptions about labour costs, we would expect the one-off direct costs to be somewhere in the region of £10,000 to £150,000.[873] In addition to the one-off direct costs, we expect this type of measure to require ongoing maintenance costs to ensure the functionality continues to operate as intended. We assume this would be 25% of the one-off costs and so we would expect it to be approximately £2,500 to £37,500 per year.

21.87 The direct cost to implement these features are likely to be dependent on the complexity of the service's system, the nature of how users typically interact on a service and the extent of organisational overheads required to implement changes. These are likely to vary significantly across services as they are influenced by the design of the service, and are likely to increase for larger services, which tend to be more complex. In some circumstances, there may also be some cost synergies with the implementation of Measure 7A as proposed in our Illegal Harms Consultation (Non-connected accounts do not have the ability to send direct messages to children using a service).

21.88 Some service providers implementing this measure would also be implementing the equivalent Measure 9A in our Illegal Harms Consultation (see 'Interaction with Illegal Harms' above). For those service providers already implementing the equivalent Illegal Harms measure we consider there are no or only negligible costs resulting from this measure because they are substantively the same feature change, and for these services the benefits would already extend to protection against certain kinds of content harmful to children. The costs (and additional benefits) for this measure are incurred where services are implementing this measure, but not the same proposed Illegal Harms measure.

Potential indirect costs

21.89 For some services, global blocking of all non-connected users (with Measure US2b) could fundamentally alter the ways in which users of the service interact. They may be less likely to interact with other unknown users, which could reduce engagement and use of a service. Depending on the business model of the service,[874] a reduction in usage could lead to a reduction in engagement with the overall service, which could in turn reduce revenue.

21.90 Interaction between user accounts differs across different U2U services, according to the functionalities that are employed. For instance, there may be some services where user connections and communication are more central to the service's functioning, and in others these features may be added value features to the wider service functionalities. This can lead to considerable variation in the indirect costs across different services.

21.91 However, any impact on engagement and usage rates is difficult to predict. While the overall effect on engagement may be negative for some users, there may be a countervailing positive impact to other users of a service which could help mitigate some of this impact. For example, users may disengage with services where they encounter harmful content. If users feel safer online, they may engage more with a service, albeit potentially with fewer users. Without such measures, some users may leave a service entirely. Therefore, while overall engagement may fall for some users, other users may increase their engagement by feeling safer, limiting the overall loss of engagement.

---

[873] In this consultation we have used 2023 wage data. This results in slightly higher cost estimates in this consultation compared to our Illegal Harms Consultation for the same amount of effort. Please see Annex 12 for details.

[874] Section 7.12, Business models and commercial profiles, sets out the relationship between engagement and revenue for U2U services.

Which providers we propose should implement this measure

21.92 The presence of certain user interaction features on U2U services can present a risk to children from bullying content, abuse and hate content and violent content. This measure provides children with key tools of the ability to block or mute other users which can help to protect them from encountering such harmful content. Given that the measure targets a specific risk factor (user interaction functionalities), we believe it can deliver distinct incremental benefits over and above other measures we propose in our Codes, which are more general in nature, or which target different specific risk factors.

21.93 Services not already adopting the equivalent measure set out in our Illegal Harms Consultation (see 'Interaction with Illegal Harms' above) would incur both direct and indirect costs from implementing this measure. These could vary considerably across different types of services. Some large services would already be implementing this measure through the equivalent Illegal Harms measure, and we believe there are further incremental benefits by extending this measure to services that pose a significant risk to children from the harms specified.

21.94 We are proposing to recommend this measure to all services likely to be accessed by children with the relevant risks and functionalities. As noted above, by targeting a specific risk factor, the measure can materially improve children's safety incrementally to other measures.

21.95 We recognise that costs are material, and this could mean that they are hard to bear for services who are smaller. While the cost of implementing this measure is likely to scale to some degree with the size of the service, we understand that in some cases this measure could require significant redesign of systems. Therefore, we cannot be confident that the costs to small services would be at the low end of our estimated range in the 'Impacts to services' sub-section. Because of this, we recognise the possibility that a minority of small businesses implementing this measure could struggle to carry this cost which could discourage some services from serving UK users, or discourage entry to the UK market. This could harm users who benefit from accessing these services.

21.96 In addition, we recognise the potential indirect impact on services where these measures, particularly global blocking, impacts the use or functioning of a service which could lead to lower engagement with a service. However, we consider that this is likely to be counteracted to some extent by a positive engagement effect where users are less likely to encounter harmful content on the service.

21.97 Overall, we consider that where the relevant user interaction functionalities are present and there is a risk of harm to children associated with these functionalities, that this risk should be reduced through the introduction of blocking and muting features. Given the significant benefits of protecting children from bullying content, abuse and hate content and violent content, we consider that it is proportionate to recommend that all services at risk of these kinds of content and that offer the relevant user interaction functionalities implement this measure.

21.98 We have provisionally concluded to apply this measure to all U2U services likely to be accessed by children that are medium or high risk for bullying content or abuse and hate

content or violent content, enable users to interact by means of user profiles,[875] and have at least one of the following functionalities: User connections; posting content; and user communication more generally (including but not limited to direct messaging and commenting on content).[876]

Provisional conclusion

21.99 Given the harms this measure seeks to mitigate in respect of bullying content, abuse and hate content and violent content, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU G1 in Annex A7.

Measure US3: Provide children with the option of disabling comments on their own posts

Explanation of the measure

21.100 Unlike private messages, comments on posts are visible to any users who have access to that content, for example where they are connected to the user or it is a public account. This includes any comments that contain harmful content.

21.101 We are proposing to recommend that services implementing this measure offer every child user the option of preventing any users from commenting on content they have posted. It would also allow users to reactively disable the comments section after upload so that all comments disappear if, for example, a post has attracted harmful comments. This measure would only apply to services that offer a comments functionality.

21.102 These measures can help to protect children from encountering bullying content, abuse and hate content, and violent content perpetrated either by a specific person, or by other users in general, where they think they are at risk of being targeted because of one of the characteristics specified in the Act (race, religion, sex, sexual orientation, disability and gender reassignment).[40]

21.103 As discussed in 'Interaction with Illegal Harms' above, this measure mirrors Measure 9B ('Measure to give users the ability to disable comments') in our Illegal Harms Consultation. Some large service providers implementing this measure would also be implementing the equivalent proposed measure in our Illegal Harms Consultation.[877]

Effectiveness at addressing risks to children

21.104 Evidence has established that comment functionalities can put children at risk of bullying content, abuse and hate content, and violent content (Governance, systems and processes, Section 7.11). [878]

[875] Includes information that is displayed to other users such as images, usernames, and biographies. Characterised by users creating a user profile that shows their identity.
[876] These terms are defined in the draft Children's Register of Risk, Section 7.
[877] For details of the Illegal Harms measure and proposed segmentation see Protecting people from illegal harms online: Volume 4, Section 16, Measure 1, page 281.
[878] Section 7.5, Bullying content; Section 7.4, Abuse and hate content; Section 7.6, Violent content.

21.105 By disabling comments on their own content, children are less likely to be exposed to bullying content, abuse and hate content, and violent content. Ofcom research details that disabling comments is one of the features children discussed to limit how others interact with them and their posts online, and one they said is important for mitigating bullying content in particular.[879]

21.106 We understand various large U2U services have already implemented measures to give users greater control of the comment functionality.

21.107 Examples of large U2U services that have implemented this type of measure are:

- **Instagram** enables users to disable all comments or block certain users from commenting.[880] It also allows for comment filters to be applied to filter certain words from appearing in comments on posts.[881]
- **X** allows users to restrict replies to tweets by allowing to comment only people the user follows or mentions.[882]
- **Facebook** allows users to choose who can comment on uploaded posts, giving users the choice between 'public', 'friends and established followers', 'friends' or 'profiles and pages you mention'.[883] Users can also block comments from specified users, and filter comments, with the option to "hide offensive comments" or manually filter out key words. It also allows users to disable the comment functionality on Facebook Live videos for all users, or to restrict to followers, comments with over 100 characters, comments from accounts that are over two weeks old, and comments from accounts that have followed the content creator for at least 15 minutes.[884]
- **TikTok** allows users to disable comments on their videos, as well as setting rules around who can comment based on their connection. Settings for under-16 users are set to 'friends only' for comments by default. It also has various comment filters including keyword filters.[885]
- **YouTube** gives users the option to disable comments on videos at any point after the video has been uploaded, as well as blocking certain accounts from commenting. It also allows for comment disabling on livestreams.[886]

21.108 We note that some services offer users a range of comment control tools. These are beyond the options we have considered here. While we are supportive of these tools as a means of empowering users to exercise more control over comment functionalities, at this stage we have limited evidence around more granular controls, and have concerns given the risk of unintended consequences with regard to uneven impacts on freedom of expression and likely higher implementation costs.

21.109 We recognise that giving users the ability to disable comments could result in a negative impact when users are unable to reply to posted content. For example, users would not have the ability to comment in a supportive way. However, we think that the benefits of this

---

[879] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.
[880] Meta, 2021. Introducing new tools to protect our community from abuse. [accessed 10 April 2024].
[881] Hootsuite (Hirose, A.), 2022. How to Manage Instagram Comments. [accessed 10 April 2024].
[882] TechCrunch, 2020. Twitter now lets everyone limit replies to their tweets | TechCrunch [accessed 10 April 2024].
[883] Meta, Facebook Help Centre. Commenting. [accessed 10 April 2024].
[884] Nerds Chalk, 2021. How To Turn off Comments on Facebook Live. [accessed 10 April 2024].
[885] TikTok, Commenting. [accessed 10 April 2024].
[886] Sprout Social, 2022. YouTube Comments: A Complete Guide. [accessed 10 April 2024].

measure are likely to outweigh any potential negative impacts associated with restricting the ability of others to comment.

21.110 This measure would provide specific additional protections for children in the case of services that are implementing this measure but not implementing the equivalent measure for Illegal Harms. For those services implementing both measures we consider that this would be an effective way for them to fulfil their protection of children duties in relation to the above-mentioned harms, as well as their Illegal Harms duties. This is why we are recommending the measure should be included in both codes.

Rights assessment

21.111 As set out above, evidence shows that comment functionalities can put children at risk of certain kind of harmful content. Therefore, by allowing children to disable comments on their own uploaded content, children would be less likely to encounter content that is harmful to them.

Freedom of expression and association

21.112 We consider that the impacts on user's rights to freedom of expression and association would be very similar to those in relation to Measure US2 above.

21.113 We acknowledge that, if a child user chooses to disable comments on their uploaded content, this would remove an interface through which other users may receive and impart information and ideas. However, this would be a choice made solely by the user concerned and have no impact on the right of other users to express themselves freely on the service in other ways. In addition, given the risk that comment functionalities pose of exposing children to harmful content, if child users are not given the option to disable comments on their own uploaded content, this may discourage them from posting at all given the risk of encountering bullying content, abuse and hate content and violent content. We therefore consider that this measure has the potential to have positive impacts on users' right to freedom of expression as for Measure US1 and US2 above.

21.114 We therefore consider that the benefits from this measure in protecting children from harmful content would outweigh any potential negative impacts associated with restricting the ability of other users to comment on their own uploaded posts.

Privacy

21.115 As with Measure US1 above, we do not consider this proposed measure would interfere with users' rights to privacy. We also have identified similar data protection impacts as for Measure US1 above, as we expect that to the extent it is necessary for providers to process users' personal data to give effect to this, they must do so in compliance with data protection requirements. We also consider that the impacts on user's rights to privacy with regards to the use of highly effective age assurance to target this measure at child users only would be similar to those in relation to Measure US1.

Impacts on services

21.116 Below we discuss the direct costs to services from implementing the measure and potential indirect costs.

Direct costs of implementation

21.117 Services that do not currently offer the functionality to enable users to disable comments would incur one-off costs to make system changes and update the user interface. However, given this measure would only be adapting the ability to use an existing comments function, we would expect costs to be lower than introducing new features, such as the blocking and muting features outlined above. The detailed assumptions underlying our direct cost estimates are found in Annex 12.

21.118 We estimated in our Illegal Harms Consultation that the direct costs of this measure would take approximately 5 to 50 days of software engineering time, with potentially up to the same again in non-engineering time. Making assumptions about labour costs, we would expect the one-off direct costs to be somewhere in the region of £2,000 to £50,000.[887] In addition to the one-off direct costs, we expect this type of measure to require ongoing maintenance costs to ensure the functionality continues to operate as intended. We assume this would be 25% of the one-off costs and so we would expect it to be approximately £500 to £12,500 per year.

21.119 As noted above, we recognise that some large service providers implementing this measure would also be implementing the same proposed measure in our Illegal Harms Consultation.[888] For these service providers already implementing the Illegal Harms measure we consider there are no or only negligible costs resulting from this measure because they are substantively the same feature change. We also recognise that for services already doing this measure the benefits would already extend to protection against certain kinds of content harmful to children. The costs (and additional benefits) for this measure are incurred where services are implementing this measure, but not the same proposed Illegal Harms measure.

Potential indirect costs

21.120 In addition to the direct costs of implementing the function, there are also potential indirect costs that are more difficult to estimate. Indirect costs could arise if the measure leads to an increase in the disabling of comments and a decrease in user commenting in relation to (non-harmful) content. If users started to disable comments on a widespread basis, the ability of other users to interact with content would be significantly reduced. Over time this could potentially lead to lower engagement and use of the service, or even users leaving a service altogether, which could reduce revenue.[889]

21.121 However, on many services users may inherently value the comments functionality to allow commenting on their posts (and it may even be that a major purpose of users posting is to receive comments), and so we consider that it is unlikely that the measure would result in

---

[887] In this consultation we have used 2023 wage data. In this case rounding means that the quantified cost estimate as the same as that in our Illegal Harms Consultation for the same amount of effort. Please see Annex 12 for details.

[888] Protecting people from illegal harms online: Volume 4, Section 16, measure 1.

[889] Section 7.12, Business models and commercial profiles, sets out the relationship between engagement and revenue for U2U services.

the widespread removal of comments in most cases. Rather, we believe that it is more likely to be used in a targeted way, including where harmful comments occur.

21.122 We also consider that there may be a countervailing positive impact to giving users the ability to disable comments, which could mitigate some of this impact. Users may disengage with services or be less likely to post where they encounter harmful content through comments on their posts and where they cannot disable these, some users could leave the service entirely as a result. Therefore, while overall engagement may fall, some users may increase their engagement because of feeling safer, limiting the overall loss of engagement.

Which providers we propose should implement this measure

21.123 Comment functionalities on U2U services can present a risk to children from encountering bullying content, abuse and hate content and violent content. We consider that this proposed measure would be an effective means of protecting children from these risks, as it would give users the ability to proactively prevent other users from commenting on their uploads with content harmful to children. It would also allow users to reactively disable the comments section after upload if a post has attracted harmful comments.

21.124 We have provisionally concluded to apply this measure to all U2U services likely to be accessed by children that enable users to comment on content, and which are medium or high risk for at least one of: bullying content, abuse and hate content and violent content. By targeting a specific risk factor, we believe the measure can materially improve children's safety in respect of these particular harms, incrementally to other measures.

21.125 We recognise that costs are material, and this could mean that they are hard to bear for services who are smaller. In particular, we estimate a wide range of direct costs, which reflects uncertainty, and we cannot be confident that the costs to small services would be at the low end of our estimated range in the 'Impacts to services' sub-section. In addition, we recognise the potential indirect impact on services where the removal of comments lead to lower engagement with a service, but consider that this is likely to be to some extent counteracted by a positive engagement effect where users are less likely to encounter harmful content on the service.

21.126 Overall, we consider that where the relevant comment functionalities are present and there is a risk of harm to children associated with comment functionalities, we believe that this risk can be reduced through the introduction of this proposed measure. Given the significant benefits of protecting children from bullying content, abuse and hate content and violent content, we consider that it is proportional to recommend that all services at risk of these kinds of content and that have comment functionalities implement this measure.

Provisional conclusion

21.127 Given the harms this measure seeks to mitigate in respect of bullying content, abuse and hate content and violent content, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCU G2 in Annex A7.

Measure US4: Provide information to child users when they restrict interactions with other accounts or content

Explanation of the measure

21.128 In our Illegal Harms Consultation, we set out evidence that suggests that children often do not report to services following harmful interactions on the platform, such as unwanted sexual messages or online grooming. This implies that there would be significant benefits if child users are given more robust and accessible information to make informed choices about reporting on the service.

21.129 In our Illegal Harms Consultation, we proposed draft Measure 7B which seeks to address this by providing information to children when they are taking action against another user. We consider that an equivalent measure in our Children's Safety Codes would protect children more broadly in relation to content that may be harmful to them, by increasing children's awareness and understanding of the functionalities available to restrict such harmful content and prompting further action to mitigate the risk of encountering other harmful content.

21.130 We therefore propose to recommend that services implementing this measure provide children with information when they take restrictive action against another account or content, to support them to increase their safety and provide information on the effect of the action taken.

21.131 Functionalities that 'restrict interaction' include blocking, muting and content restriction tools. Refer to 'Definition Box 1: Glossary of key functionalities', in the Introduction of this section for a definition of content restriction tools. Services may use different names for these functionalities.

21.132 The information should include but is not limited to:

21.133 Information on the effect the action taken will have on interactions with the account or content in question. For example, an explanation of the kind of content or functionalities that a user may be restricting (such as future encounters with similar content) and where applicable, confirming whether the user will be made aware of the fact that the child has taken action against them.

21.134 Information to support child users to increase their safety on the platform. For example, information on further steps the child user can take to limit interactions with the account or further restrict content, hyperlinks to security settings, supportive knowledge around online safety, a prompt to submit a report, or, if applicable, age-appropriate support materials as explained in 'Measure US5: Signpost to children to support at key points in their user journey' in this section.

21.135 We are not making specific recommendations about how the information should be presented and are encouraging services to establish their own practices in consideration of their service type and user base. However, the information should be easy for child users to understand and displayed prominently to them as soon as possible after they take restrictive action against a user or piece of content.

21.136 We will consider at a later date whether to extend the Illegal Harms supportive information measure to include that information should be provided when a child user takes action against content, in addition to users, ahead of the Illegal Harms Statement being published.

Effectiveness at addressing risks to children

21.137 In our Illegal Harms Consultation, we set out evidence that timing and relevance of such interventions are particularly important in achieving desired effects.[890] We consider that providing relevant supportive information at the point a child user has taken action to restrict content and/or limit a user interaction would be effective in increasing a child's awareness of relevant restriction tools and in helping them make more informed choices on how to further restrict harmful content or interactions either at that point or in the future.

21.138 Evidence in Governance, systems and processes, Section 7.11 suggests that, instead of reporting, many children use functionalities such as blocking users and content restriction tools to protect themselves from harmful content. Yet, further evidence also suggests that children are less aware of blocking and reporting functions on services, and therefore may benefit from supportive information at the point of restricting interaction with either a user or content to make informed choices.[891]

21.139 Evidence shows that information on methods to further restrict content or user interaction can be effective in influencing users to report, and that prompts can influence people to make safer choices.[892] Ofcom behavioural research trials also showed that when we increased the visibility of a reporting tool and prompted users to report content if they disliked or commented on a video, 11% of adult users reported at least one video, compared to 4% when we increased the visibility of the reporting tool only, and 1% in the control arm. There was no evidence of an increase in over-reporting, nor an increase in inaccurate reports.[893]

21.140 The above evidence suggests that providing users further information at an appropriate time on how to restrict interaction, such as reporting, can be an effective way to increase reporting by users without reducing the accuracy of reports. Although this research was conducted with adults, based upon the findings it is reasonable to infer that similar information for children can be used to encourage child users to adopt better privacy practices on social media.[894]

---

[890] Volume 4 (ofcom.org.uk) paragraph 18.107. [accessed 22 March 2024].
[891] Evidence about mixed awareness of safety features and low reporting levels among children can be found in Ofcom, 2023. Online Nation.
[892] For example: European Commission, 2019. Study on media literacy and online empowerment issues raised by algorithm-driven media services. [accessed 21 September 2023]; US Food and Drug Administration, 2019. Communicating Risks and Benefits: An Evidence-Based User's Guide. [accessed 22 March 2024]; Tussyadiah I., Miller G., Li S. and Weick M., 2021. Privacy nudges for disclosure of personal information: A systematic literature review and meta-analysis. PLoS One, 16 (8). [accessed September 21 2023]; Acquisti et al., 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. ACM Computing Surveys, 50 (3). [accessed 22 March 2024].
[893] Ofcom, reissued 2023. Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour, pages 35-36. [accessed 22 April 2024].
[894] For an example of a study exploring this theme with teenagers, see Alemany, J., del Val E., Alberola, J., García-Fornes, A., 2019. Enhancing the privacy risk awareness of teenagers in online social networks through soft-paternalism mechanisms. *International Journal of Human-Computer Studies.* 129. [accessed 4 September 2023].

21.141 Ofcom research into children's experiences of violent content online, found that child users felt reporting would not have any impact, this was a key barrier in discouraging children from reporting violent content.[895] By providing children with further information on the effect their report will have on their online interactions with the content or user, and other safety information, this measure could lead to children feeling a greater sense of action from the service in response to their report or restrictive action. In addition, this may also increase the chances of children continuing to use restrictive actions if they perceive the service is having a positive impact on their online experiences.

21.142 Some services already offer measures that provide information on how users can further restrict content or a user.[896] For example:

21.143 **WhatsApp** told us that they have in-app tools that prompt users to restrict unwanted interactions. For example, if a user receives a message from someone who is not saved in their WhatsApp contact list, the service immediately asks if the user would like to "block" or "report" that other user.[897]

21.144 **Pinterest** said that if a user declines a message request from a user outside of their network, they are presented with the option to block or report that person. The contact request also warns users not to share confidential information.[898]

21.145 In response to our 2022 Illegal Harms Call for Evidence (our 2022 CFE), the ICO pointed to Standard 13 of their Children's code, which envisages that nudge techniques can be used for pro-privacy reasons and 'suggests that services should consider nudging to promote the health and wellbeing of child users'.[899]

21.146 While such tools may be effective for preventing an individual child from encountering content harmful to children, they may not make the service aware of that content. This means that services would be less likely to review it for being content harmful to children and take steps to protect other children from it. User reports are a key mechanism for users to bring content harmful to children to services' attention, particularly on smaller services which may not use proactive detection methods. We therefore consider that providing information to children about their options to further limit interaction with users or content that is harmful to children, may encourage children to report more of this content that could play an important role in protecting them from it.

What should the information say and look like?

Presentation of the information:

21.147 Evidence suggests that various factors including length, colour, and language can contribute to the effectiveness of information measures. For example the ICO's Children's code provides guidance on the interests, needs and evolving capacity of children at different ages.[900]

---

[895] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children. [accessed 22 April 2024]
[896] Meta response to 2022 Illegal Harms Call for Evidence, Q19, page 36; and Roblox response to 2022 Illegal Harms Call for Evidence, Q5, page 4.
[897] WhatsApp's response to 2023 Protection of Children Call for Evidence, Q16,17,18, page 12.
[898] Pinterest's response to 2023 Protection of Children Call for Evidence, page 6.
[899] ICO response to 2022 Illegal Harms Call for Evidence, page 5.
[900] See ICO, 2020. Age appropriate design: a code of practice for online services [accessed 16 April 2024]

21.148 While some research indicates that prompts can be effective, other research suggests that they can be perceived as annoying, and that excessive frequency could lead to alert fatigue where people do not engage with the information.[901] However, we do not consider this means they are necessarily ineffective in meeting the desired objective under this proposal given some children use existing blocking functions to avoid further exposure to harmful content (Governance, systems and processes, Section 7.11). In addition, we consider the intention is that relevant information is provided to a child user at a specific intervention point rather than on a frequent or repetitive basis so child users are less likely to be subjected to alert fatigue.

21.149 There is also some variation in the way services present information measures to users. As an example, some platforms present such information as a pop-up, while others embed it within a user interface including information banners or support buttons.

21.150 The current evidence does not support a single 'best practice' approach. It is for services to adapt, test and evaluate the effectiveness of the measure to their individual service. We are not therefore proposing to make specific recommendations around how the information should be presented.

Content:

21.151 The exact wording of the provisional information when a child user is restricting content or another account will depend on the individual service's functionalities. We do, however, expect that the information is clear, comprehensible and easy for a child user to understand and is displayed prominently to them at the relevant critical point.

21.152 While we are not providing detailed guidance on content, the information presented to children should include (but is not limited to):

- What actions they have restricted. For example, clarity on if the child user has blocked one piece of content or all similar pieces of content on the service and how this will affect their future on platform experiences.
- Information on further actions a child user can take to increase their general safety on the platform. This could include further actions to restrict their interaction with the user or content such as links to reporting channels, signposting to online resources such as user guides or terms of service and prompts to review security and privacy settings.

Rights assessment

21.153 The proposed measure recommends services provide supportive information to users when taking restrictive action on content or a user, which would help inform the user of the effects of the action they have just taken and make them aware of options to take further restrictive action (for example, blocking, muting or reporting) and additional safety information. By providing supportive information the service would empower users to make informed choices on taking action to prevent and protect children from encountering harmful content, which would mitigate the risks and impact of such content, in line with the legitimate aims of the Act.

---

[901] Micallef, N., Just, M., Baillie, L., and Alharby, M., 2017. Stop annoying me!: an empirical investigation of the usability of app privacy notifications. Association for Computing Machinery. Proceedings of the 29th Australian Conference on Computer-Human Interaction. [accessed 22 March 2024].

21.154 As with Measure US1, we consider below the potential impacts on users' rights to freedom of expression and association and privacy. As with Measure US1, services may apply this measure to child users only where they use highly effective age assurance to identify child users, or else would need to apply this measure to all users (i.e. including adult users), and we have therefore assessed the potential impact under both scenarios. We also consider that where services decide to apply this measure to all users, such supportive information could also be beneficial for adults who may not be aware of the relevant tools available to restrict content and limit interactions.

Freedom of expression and association

21.155 We consider that the impacts on user's rights to freedom of expression and association would be very similar to those in relation to Measures US1 and US2 above. The proposed measure would require the service provider to present information to prompt possible further restrictions of content that is harmful to children, or to further limit interactions with a creator of content. This could interfere with a user's rights, as the prompt may dissuade users from accessing and receiving information or may dissuade them from connecting with other users, and would potentially create additional frictions in their online experience. However, we consider that any impact on users' or services' rights to freedom of expression and association is minimal in that any such restriction would follow only from the informed choice of the user concerned, in the event they decide to take action to further limit interaction with content or other users. In addition, the design of the measure seeks to limit frictions to the user experience or the risk of a user inadvertently taking further action as a result of a prompt, as we propose to stipulate that the information presented should be clear, comprehensible and easy for a child user to understand and displayed prominently to them at the relevant critical point.

21.156 As for Measure US1 above, we also consider this measure could have a positive impact on users' rights to freedom of expression and freedom of association, for example, children may feel more able to share and impart ideas and information where they feel safe online as they would receive appropriate information when they decide to restrict their interaction with content or with other users.

21.157 Having carefully considered the potential impact on users' and service providers' rights to freedom of expression and association, for the reasons set out above, our provisional conclusion is that the limited degree of interference with these rights would be proportionate given the substantial public interest that arises in the protection of children.

Privacy

21.158 As with Measure US1 above, we do not consider this proposed measure would interfere with users' rights to privacy, and indeed might have positive benefits for children's rights to privacy in that it would give them additional options for deciding how to share their personal information and content online. We also have identified similar data protection impacts as for Measure US1 above, as we expect that to the extent it is necessary for providers to process users' personal data to give effect to this, they must do so in compliance with data protection requirements. We also consider that the impacts on user's rights to privacy with regards to the use of highly effective age assurance to target this measure at child users only would be similar to those in relation to Measure US1.

Impacts on services

21.159 Below we discuss the direct costs to services from implementing the measure and potential indirect costs.

Direct costs of implementation

21.160 This measure applies where services have certain functionalities that allow users to take restrictive action against another account or content. Costs will be higher where services have both of these types of functionalities, and lower where they only have one of the two. Our estimates assume that services have both types of these functionalities. Direct costs for services that do not already meet this measure are likely to be largely one-off costs.

21.161 Costs would include developing information to present to users. We expect these costs to be relatively small, as it would largely require providing simple explanations of relevant features and providing links to other existing materials.

21.162 The larger share of one-off costs would relate to implementing system changes to trigger the provision of information at the point where a user restricts interaction with an account or content. We are not proposing to prescribe how exactly services should provide information. As a result, costs may vary according to the approach a service takes.

21.163 The steps required to implement this measure will vary based on the design of the service, and how many functionalities a service has that restrict action against another account or content. The cost and time to implement will also vary based on the complexity of the design of the measure. In estimating indicative costs, we have anticipated that implementation for each functionality may involve feature design, including user journey mapping and changes to the UI/UX; and technical development to integrate the provision of supportive information with existing workflows which will include changes to the backend infrastructure.

21.164 Technical development costs could be material if services do not already have a relevant system to provide user prompts. They could, however, be much lower if they have an existing system to provide warnings or prompts in other contexts. Services may also incur costs associated with testing and evaluating the format and delivery methods, with the possibility of changing these if they are not working well. The level of these testing costs is likely to depend on complexity of the service, and the extent to which the service evaluates effectiveness or monitors impact on user behaviour or experience.

21.165 The development steps described may need to followed multiple times depending on how many functionalities a service has that would trigger the provision of supportive information, and may require discrete sets of system changes to address action against users and against content. This process is likely to require various technical skills (such as business analysts, graphic designers, web designers, user experience or interface designers, content teams, and developers, plus quality assurance and/or testing teams). We have assumed that across all relevant functionalities, implementation could take approximately 3 to 6 months of labour time from technical occupations, for which we apply the software engineering salary category, matched with an equivalent amount of other professionals (e.g. project management.[902] We expect this could amount to costs in the region of £28,000 - £113,000.

[902] See Annex 12 for our further detail on economic assumptions and analysis.

21.166 There would also be some ongoing costs to maintain the functionalities. We assume the annual maintenance costs of this measure to be 25% of the initial implementation costs, which would be approximately £7,000 - £28,000. Ongoing running costs are likely to include regular updating of the supportive information and system maintenance costs.

21.167 We recognise that some services implementing this measure would already be implementing the supportive information measure proposed in our Illegal Harms Consultation. As discussed above, in that consultation we recommended that certain services provide information when a child user is taking action against another account. For services implementing both measures, the incremental cost of this measure would involve extending that feature to apply when a child user is taking action against content.

21.168 For such services, we expect that the incremental costs could be significantly lower than if implementing the measure from scratch. For example, if the incremental costs were around half of the costs of implementing for both action against accounts and content, then the one-off costs could be in the region of £14,000 - £56,000, with an assumed annual maintenance cost of £4,000 - £14,000. Again, the exact cost would depend on the complexity and existing functions of the system and the extent of the supportive information that is provided.

Potential indirect costs

21.169 Depending on how it is implemented, on some services users may have to choose either to follow or to ignore the prompt each time they restrict their interaction with another user or kind of content. This could alter the flow of the user experience, potentially reducing user engagement to some degree and indirectly impacting service revenue.[903] However, our measure allows services flexibility to decide how to display the information in a way that is appropriate for their service, which somewhat mitigates this risk. On balance we consider it unlikely that this measure would materially discourage a high proportion of users from using the service or discourage them from taking restrictive action against content or accounts, taking into account that the specific information provided to users under this measure is not envisaged to require a long time for users to process and respond.

21.170 We consider that this measure could lead to a higher volume of reports where a service provides information about reporting in its supportive information. While this may increase the costs of handling additional reports, it should also tend to improve children's safety online by helping the service to identify and action harmful content. Where this in turn improves user experience and engagement, it may have some indirect benefits for the service. It is possible there may also be an increase in the volume of inaccurate reports, although our behavioural research mentioned in the 'Effectiveness' sub-section above found this was not the case in an experimental setting.[904]

[903] Section 7.12, Business models and commercial profiles, sets out the relationship between engagement and revenue for U2U services.

[904] Ofcom, reissued 2023. Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour.

Which providers we propose should implement this measure?

21.171 We expect that benefits can arise from providing information to children at critical points during their user journey, including through increasing their ability to make more informed and safer choices. The measure may also increase accurate reports of harmful content, and so can enhance the effectiveness of our separate User Reporting and Complaints and Content Moderation measures, by helping services to identify and action more harmful content which would have significant benefits for children's online safety.

21.172 If implemented by large services, the measure would mean that many users, including children, can benefit from the information provided. In addition, we consider that the benefits of this measure would be greater when services are able to optimise the user experience to help mitigate the risk of alert fatigue. We believe that large services are better positioned to be able to do this as they typically have more sophisticated user interfaces and greater capacity to develop and introduce prompts, notifications or other forms of information without unduly disrupting the user experience.

21.173 The impact on children's safety from this measure is expected to be material for large services that are multi-risk. On these services there is likely to be a higher volume of content harmful to children, and we also expect that these services are more likely to have a range of relevant restriction tools. As a result, a greater number of users, including children, are likely to use tools to restrict their interaction with other accounts or content and therefore stand to benefit from understanding better how the restriction tools work and what next steps they can take.

21.174 At this stage we do not consider it proportionate to recommend this measure for large services that are not multi-risk for content harmful to children. We expect that benefits would be more limited for these services, as the relevant user or content restriction tools are likely to be used less frequently. We have also considered that, where services pose medium or high risk for a single kind of content harmful to children, there are already measures recommended in this section and others that specifically target certain risks, and the incremental benefit from this measure on such services could be more limited.

21.175 At this stage we also do not consider it proportionate to recommend this measure for smaller services. The costs of this additional measure for smaller services may be very material on top of the other measures they would already be implementing, and the risk of unintended user impacts from this measure may be higher on such services. We also believe that the incremental benefits for this measure are likely to be smaller for these services given their more limited reach, and considering that these services would in any case be in scope of other measures that will help to give children information and control over key elements of their experience online. These include the other measures in this section and User reporting and complaints Section 18. In particular, US6 in this section recommends that smaller multi-risk services provide age-appropriate user support materials covering a range of functions. As discussed further in our Combined impact assessment Section 23 we have prioritised measures for smaller services where we believe that there would be material benefits to children's safety online.

21.176 We are not proposing to recommend as part of this measure that service providers should introduce functionalities that enable users to restrict their interaction with other users or kinds of content. Rather, it would apply only to services that offer users these functionalities. Proposed Measure US2 in this section recommends that certain services should offer users options to block or mute other users.

21.177 We have provisionally concluded this measure should be recommended to all large U2U services likely to be accessed by children that are multi-risk for content harmful to children.

Provisional conclusion

21.178 Given the harms this measure seeks to mitigate in respect of content harmful to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text of this measure please see PCU E2 in Annex A7.

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

Explanation of the measure

21.179 Under the Act, providers of U2U services likely to be accessed by children have a duty to mitigate the impact of harm to children in different age groups presented by content that is harmful to children present on their services.[905]

21.180 Our evidence suggests that one way to mitigate the impact of harm posed by certain kinds of content harmful to children is by signposting children to support at key points in the user journey. We discuss this evidence further below.

21.181 We are proposing to recommend that providers of U2U services signpost children who encounter relevant kinds of content harmful to children to appropriate support at key points in the user journey.

21.182 As set out below, we are aware of evidence that suggests signposting is effective at the following three points in the user journey ('intervention points'):

1. When children report content;
2. When children post or re-post content; and
3. When children search for user-generated content on U2U services.

21.183 We also set out evidence that suggests signposting is effective at mitigating the impact of harm of the following kinds of content: suicide content; self-harm content; eating disorder content; and bullying content.

21.184 We are proposing to recommend that providers of different types of services implementing this measure should signpost children to appropriate support at each intervention point. We explain this in the 'Which providers we propose should implement this measure' section below.

21.185 As mentioned at in the 'Which users these measures apply to' section, under the measures in Section 15, Age Assurance, providers of services implementing this measure may use highly effective age assurance to apply this measure only to children or should apply the measure to all users. We discuss the impact this could have on users in the 'Rights assessment' section below.

[905] Section 12(2)(b) of the Act.

21.186 In this section we discuss the evidence for the effectiveness of this measure and how it would work in practice for each intervention point, including what is meant by 'appropriate support'. We also discuss the costs and impacts of signposting at each of these intervention points.

Harms this measure aims to mitigate

Evidence for the effectiveness of signposting to address certain risks of harm

21.187 As discussed in Section 7.2, Suicide and self-harm content, Section 7.3, Eating disorder content, and Section 7.5, Bullying content, evidence suggests that suicide, self-harm, eating disorder and bullying content can pose particularly significant risks of harm to children.[906] Academic studies also suggest that online self-help tools and support resources may be helpful for young people who have experienced suicidal feelings and other mental health concerns.[907] For example, considering the findings of these studies in the round, it seems clear that signposting to support may help to validate children's experiences and make them realise they are not alone.

21.188 The report of a 2018 cross-parliamentary inquiry into children's experiences of cyberbullying content found that 79% of survey participants aged 11 to 25 believed signposting to mental health support sites would be effective for those affected by cyberbullying content; this view was common among those who had experienced cyberbullying content and those who had not.[908] This was echoed in research commissioned by Ofcom, in which participants recommended that information shared with children should include support resources – both those being bullied and those bullying others.[909]

21.189 Participants in Ofcom commissioned research into children and young people's experiences of suicide, self-harm and eating disorder content online likewise called for providers to signpost users to support resources to mitigate the impact of those kinds of content, and counteract the large amounts of unreliable information on those topics available online.[910] In their 2022 research, 'How social media users experience self-harm and suicide content', Samaritans also called for signposting to appropriate support.[911] A report by the Royal Society for Public Health has also suggested that signposting on social media sites in

[906] Section 7.2, Suicide and self-harm content; Section 7.3, Eating disorder content; Section 7.5 Bullying content.
[907] Cohen, R., Rifkin-Zybutz, R., Moran, P., Biddle, L., 2022. Web-based support services to help prevent suicide in young people and students [accessed 15 December 2023]. Biddle, L., Derges, J., Goldsmiths, C., Donovan, J., Gunnell, D., 2020. Online help for people with suicidal thoughts provided by charities and healthcare organisations: a qualitative study of users' perceptions [accessed December 2023]. Garrido, S., Millington, C., Cheers, D., Boydell, K., Schubert, E., Meade, T., Nguyen, Q. V., 2019. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people [accessed 15 December 2023].
[908] YoungMinds, 2018. Safety Net: cyberbullying's impact on young people's mental health [accessed 15 December 2023]. 80% of those who had not been bullied felt this would be effective or very effective, compared to 79% of those who had been bullied.
[909] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.
[910] Ofcom, 2024. Experiences of children encountering online content promoting eating disorders, self-harm and suicide.
[911] Samaritans 2022. How social media users experience self-harm and suicide content [accessed 15 December 2023].

particular can be an effective way to encourage young people to engage with healthcare services.[912]

21.190 In response to our 2023 CFE, the NSPCC, Papyrus, Ygam, SWGfL, UKSIC, Nexus and Refuge recommended signposting children exposed to harmful online content to support resources.[913]

21.191 This evidence suggests that both children and experts in children's mental health and wellbeing recognise the value of timely and appropriate signposting to support for children exposed to suicide, self-harm, eating disorder or bullying content. There is less evidence for the effectiveness of signposting to support for exposure to other kinds of content harmful to children.

21.192 For these reasons, we are proposing that providers should signpost children to support at key points in the user journey when they encounter suicide, self-harm, eating disorder or bullying content. We may consider whether to extend this measure to other kinds of content harmful to children as part of our future work.

Explanation of 'appropriate support'

Evidence regarding support appropriate for children

21.193 We know that some providers already have a variety of signposting resources in place, for example written and audio-visual material, interactive on-platform chatbots, and helplines. We think providers are best placed to tailor those measures to the needs of children using their services. For this reason, our proposed measure does not specify the format of support resources or whether providers should produce their own support resources or signpost to support provided by third parties. We would encourage providers to have regard to research which suggests some children and young people want clear, brief, on-platform support for mental health, and can prefer text-based interventions, such as direct messaging to verbal communication when seeking help.[914]

21.194 We are aware of evidence that where providers already signpost users to support, this support is not always appropriate to their needs. The Samaritans research mentioned above found that 53% of survey participants said that the support they were directed to was not relevant to them. Focus group participants described signposting as generic, often with details of helplines outside the UK.[915] Our research into children and young people's experiences of suicide, self-harm and eating disorder content found that support numbers and contacts were sometimes based in other countries. Participants called for support

[912] Royal Society for Public Heath, 2017. #Status Of Mind: Social media and young people's wellbeing and mental health [accessed 15 December 2023].
[913] NSPCC response to 2023 Protection of Children Call for Evidence. Papyrus response to 2023 Protection of Children Call for Evidence. Ygam response to 2023 Protection of Children Call for Evidence. SWGfL response to 2023 Protection of Children Call for Evidence. UKSIC response to 2023 Protection of Children Call for Evidence. Nexus response to 2023 Protection of Children Call for Evidence. Refuge response to 2023 Protection of Children Call for Evidence.
[914] Cohen, R., Rifkin-Zybutz, R., Moran, P., Biddle, L., 2022. Web-based support services to help prevent suicide in young people and students [accessed 15 December 2023].
[915] Samaritans 2022. How social media users experience self-harm and suicide content [accessed 15 December 2023].

resources to be relevant, appropriate for the region in which the user was based and produced by authoritative sources, such as the NHS.[916]

21.195 This reflects similar findings in a number of other studies, which indicate that support resources are not always appropriate to the users signposted to them. Evidence suggests that children may be particularly discouraged by information presented to them that is not appropriate for their age.[917] One study looking at how effective digital mental health interventions are for treating young people with depression found evidence that interventions which seemed to be designed for much younger ages put children off.[918]

21.196 Given the importance of support resources being relevant, appropriate, accessible to children in the UK and authoritative, we propose to set out some high-level principles regarding the nature of the support to which children should be signposted.

'Appropriate support' principles

21.197 'Appropriate support' could take any format that meets the principles set out below. This might include but is not limited to written or audio-visual materials, interactive on-platform chatbots, or helpline numbers.

21.198 'Appropriate support' is support that is:

a) Relevant to the specific kind of content in question and the way children are affected by it;
b) Comprehensible and suitable in tone and content for children;
c) Accessible to/can be accessed by children in the UK; and
d) Produced in consultation with an expert third-party organisation, if the provider wishes to produce its own support; or
e) Produced by an expert third-party organisation, if the provider wishes to signpost to third-party resources.

21.199 Expert third party organisations are those that meet the following criteria:

a) Have expertise in the relevant harm; and
b) Have support resources appropriate for children and, if the service is signposting to support for individuals, are able to appropriately support children in the UK.

21.200 We think these criteria are the minimum necessary to ensure support resources are relevant, appropriate, accessible to children in the UK and authoritative.

21.201 We do not consider it proportionate to recommend that providers should target different support to different age users, as this is likely to be difficult and costly to do accurately (although providers may choose to do this if they wish). Rather, providers should ensure the support they signpost children to is comprehensible and suitable in tone and content for the youngest person permitted to use the service without permission from a parent or guardian.

---

[916] Ofcom, 2024. Experiences of children encountering online content promoting eating disorders, self-harm and suicide.
[917] Cohen, R., Rifkin-Zybutz, R., Moran, P., Biddle, L., 2022. Web-based support services to help prevent suicide in young people and students [accessed 15 December 2023]. Biddle, L., Derges, J., Goldsmiths, C., Donovan, J., Gunnell, D., 2020. Online help for people with suicidal thoughts provided by charities and healthcare organisations: a qualitative study of users' perceptions [accessed December 2023].
[918] Garrido, S., Millington, C., Cheers, D., Boydell, K., Schubert, E., Meade, T., Nguyen, Q. V., 2019. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people [accessed 15 December 2023].

21.202 We are aware of a number of services that currently signpost users in the UK to UK-specific support. For example, in response to our 2023 CFE, X told us that the support they provide depends on the user's location, and in the UK they have partnered with Samaritans.[919] Snapchat likewise prompts potentially at-risk users to resources provided by local partners.[920]

21.203 A number of organisations, such as specialist charities, that currently work with service providers to develop and/or provide resources relating to suicide, self-harm, eating disorder and bullying content. We understand they work with service providers both on an individual basis, and as part of larger forums. For example, in response to our 2023 CFE, Samaritans told us that they have developed an 'Online Excellence Programme' which includes 'industry guidelines for responding to self-harm and suicide content [and] an advisory service for sites and platforms'.[921]

21.204 However, we are conscious that there may be a very large number of services implementing these recommendations, some with very large user bases. Many of these providers may not already be signposting to support resources, meaning this measure could lead to a significant increase in the number of children being signposted. While this is the intended outcome of the measure, we are conscious this could also lead to third-party organisations providing one-to-one support becoming overwhelmed by requests. We therefore propose to recommend that if a provider wishes to signpost directly to support services or helplines run by organisations that offer support to individuals, they should obtain permission from that organisation to do so, unless the organisation is in the public sector (e.g. NHS or Department of Education).[922] The NSPCC recommended this in their response to our 2023 CFE.[923]

Intervention point 1: signpost children when they report specific kinds of content harmful to children

Evidence for signposting at this intervention point

21.205 Given low rates of reporting among children,[924] we think it is reasonable to assume that the majority of children who report content (including suicide, self-harm, eating disorder and bullying content) do so because they find it upsetting or distressing rather than for other reasons, such as finding the content annoying. Children participating in our 2024 research into children's attitudes to reporting indicated that they were more likely to take action

[919] X (formerly known as Twitter) response to 2023 Protection of Children Call for Evidence.
[920] Ofcom, 2022. Ofcom's first year of video-sharing platform regulation.
[921] Samaritans response to 2023 Protection of Children Call for Evidence.
[922] This measure is different from Measure RS2 in Section 20, which requires search services to provide crisis support to users in response to suicide, self-harm and eating disorder search requests. This is because that measure applies only to large general search services, of which there are very few, and which we understand already signpost to support in a number of instances. Unlike for U2U services, we are not aware of concerns among third-party organisations that the crisis support measure for search services could lead to their support services becoming overwhelmed. We are therefore not proposing that search service providers should seek permission to signpost to third-party crisis support materials.
[923] NSPCC response to 2023 Protection of Children Call for Evidence.
[924] Measure US4 above aims to increase accurate reporting by children. However, reporting by children is currently so low, that even with this potential increase we still consider the majority of children who report suicide, self-harm, eating disorder or bullying content are likely to do so because they have been distressed by it.

against content, for example, by reporting or blocking it, when they thought the content posed a more severe risk of harm.[925]

21.206 Our Online Experiences Tracker 2023 found that only 3% of 13–17-year-olds complained to the service about the most recent piece of potentially harmful content they encountered online. 10% of 13–17-year-olds clicked the report or flag button or marked the content as junk. The more 'bothered or offended' 13-17-year-olds were, the more likely they were to report or flag content. For example, of the 13-17-year-olds who had recently seen harmful content, 8% who were 'not at all bothered or offended' reported the content. This rose to 19% of those who were 'slightly bothered or offended' and 38% of those who were 'really bothered or offended'.[926] We therefore consider that signposting children who report to appropriate support, is likely to capture mainly children who have been negatively affected by that content.

21.207 Participants in our research into children's and young people's experiences of suicide, self-harm and eating disorders called for providers to signpost to support when children report or block such content.[927] Childrens experts interviewed for our research into children's experiences of bullying content also called for more immediate links to emotional support alongside reporting mechanisms. Children who participated in the research echoed this view, recommending providers should immediately signpost children to support when they report bullying content.[928]

21.208 This evidence suggests that submission of a report of harmful content is an effective intervention point at which to signpost children to support.

Explanation of the measure in practice

21.209 Evidence suggests that children and young people find immediate signposting to support particularly valuable.[929] Given the potentially severe consequences for children of encountering such content, we consider the value of signposting to support is likely to increase the sooner it takes place. We therefore propose to recommend that providers should signpost children to support as quickly as possible following a report being submitted. In its simplest form, this could be done in an automated acknowledgement of the report. In most instances we expect signposting would occur within a few seconds, although we recognise there may occasionally be circumstances that mean longer delays are unavoidable.

21.210 Providers of services implementing this measure can be split into two groups: those that already have methods in place to identify what kind of content a child is reporting at the point when the report is submitted and those who do not. We propose that each group should implement this measure slightly differently. We discuss each group in turn below.

---

[925] Ofcom, 2024. Children's Attitudes to Reporting Content Online.
[926] Ofcom, 2023. Online Experiences Tracker. Findings are derived from analysis of raw data, about respondents aged 13-17: their actions in response to their most recent harmful experience (Q15) against how impacted they were by their most recent harmful experience (Q14b), for any named harm. (% of respondents aged 13-17, across reported impact levels, taking any given action). Base sizes for the 3 groups are 237 (not at all bothered/offended), 302 (slightly bothered or offended) and 95 (really bothered or offended).
[927] Ofcom, 2024. Experiences of children encountering online content promoting eating disorders, self-harm and suicide.
[928] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.
[929] Cohen, R., Rifkin-Zybutz, R., Moran, P., Biddle, L., 2022. Web-based support services to help prevent suicide in young people and students [accessed 15 December 2023].

Providers who already have methods to identify the kind of content being reported when the report is submitted

21.211 Many providers already have methods in place to identify the kind of content a user is reporting at the time a report is submitted.

21.212 We know many providers do this by asking users to categorise the content they are reporting as part of the reporting process. For example, YouTube, Pinterest, Snapchat and TikTok all ask users to categorise the content they report.[930] Although these services offer users different categories, they all include some or all of suicide, self-harm, eating disorders or bullying.

21.213 Providers that have categories but do not currently include all of these kinds of content should consider whether it would be appropriate to the risks posed by their service to include those categories, in order to enable them to identify when children are reporting these kinds of content and signpost them to appropriate support.

21.214 We know some providers use automated content moderation tools. These may also allow providers to identify the likely kind of content a user is reporting at the time when a report is submitted, before it undergoes further moderation.

21.215 Where providers already have methods in place that enable them to identify the kind of content being reported, they can target their signposting to children reporting suicide, self-harm, eating disorder or bullying content only.

21.216 Whatever method providers use to identify the kind of content being reported, there is a risk that it may lead to content being misidentified. Where providers ask users to categorise the content they are reporting, there is a risk that they may do so incorrectly. This is a particular risk for children, who may not always understand the categories they are presented with if these are not designed with children in mind. Where providers use automated content moderation tools, these may also sometimes classify content incorrectly.

21.217 Misidentification of the kind of content being reported could mean that some children are signposted to support that is not appropriate for the kind of content they reported. It could also mean that some children in need of support may not receive it.

21.218 We consider that Measure UR2 in Section 18 would help to mitigate the risk of users miscategorising the content they report by recommending that all information and processes relating to complaints should be accessible and comprehensible. This means that providers would need to ensure the categories they present to users during the reporting process are comprehensible to children.

21.219 We think that Measure CM3 in Section 16 would go some way towards mitigating the risk of automated content moderation technology misidentifying the likely kind of content being reported by recommending that providers of large services and services that are multi-risk for content harmful to children should set performance targets for the accuracy of their content moderation processes, including any automated content moderation technologies they use, and ensure they are well resourced in order to meet those targets.

21.220 Despite the risk that some children may not be signposted correctly, we consider that providers that already have methods in place to identify the kind of content being reported

[930] Google response to 2022 Illegal Harms Call for Evidence. Pinterest response to 2023 Protection of Children Call for Evidence. Ofcom, 2022. Ofcom's first year of video-sharing platform regulation.

should be able to target signposting to children reporting suicide, self-harm, eating disorder or bullying content only. This is because we consider the benefits of targeting signposting to the kind of content being reported outweigh the risks of signposting incorrectly.

Providers who do not already have methods to identify the kind of content being reported when the report is submitted

21.221 Where providers do not currently have methods in place to identify the kind of content being reported, we are not proposing they should introduce these for the purpose of this measure. We set out our reasons for not recommending providers ask users to categorise content when reporting in Section 18, User reporting and complaints. We set out our consideration with respect to automated content moderation in Section 13: Overview of Codes.

21.222 Where a provider does not have such a method in place, we consider that in order to ensure children exposed to suicide, self-harm, eating disorder or bullying content on their service can benefit from being signposted to support, they should signpost all children to support for each of these kinds of content immediately following reporting. This should mean that when children report one of those kinds of content, they are still signposted to relevant support, even if it is presented alongside support that is not relevant to them.

21.223 We recognise that signposting in this way is less targeted. However, we consider that the risks of signposting children to irrelevant support are low. The 'appropriate support' principles set out above recommend that support should be appropriate for children to use, meaning that there should not be any risk of harm posed by signposting, even to those who do not need it. We do not consider that untargeted signposting need add additional friction to the user journey, since there are several ways providers could avoid this, for example by including links to support in an acknowledgement of the report (refer to UR3 for the measure on acknowledgement of receipt of complaints). There is a risk that signposting to all children who report could contribute to alert fatigue, leading children to engage less with information provided to them. However, we have not seen evidence that suggests this is a significant risk. On balance, we therefore consider the benefits of signposting outweigh the risks of less targeted signposting.

Intervention point 2: signpost children when they post or re-post specific kinds of content harmful to children

Evidence for signposting at this intervention point

21.224 Evidence suggests that children who create and re-post suicide, self-harm, eating disorder or bullying content would also benefit from being signposted to support.

21.225 Participants in our research into children's experiences of bullying content explicitly called for those who were bullying others to be signposted to support.[931] Children who post or re-post bullying content may be engaging in (and possibly also be victims of)[932] bullying behaviour or they may not understand the risk posed to others by such content. In either case, they are likely to benefit from being signposted to support, whether to help them with their own concerns or to understand the impact of their actions on others.

[931] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK.
[932] Evidence in Section 7.5, Bullying content, suggests that many children who perpetrate bullying have themselves been the target of bullying.

21.226 Participants in our research into children's experiences of suicide, self-harm and eating disorder content suggested that some children post self-harm content asking for informal support from their community and peers.[933] Participants shared that the intention of people who posted recovery content was not always clear and suggested that some people posting 'recovery content' relating to suicide, self-harm and eating disorders do so to get attention in the form of likes, comments and followers.[934] This may also be indicative of a desire for support. This suggests that signposting children who post or re-post such content has the potential to reach some of those who would benefit most from accessing authoritative sources of support.

21.227 We understand that signposting at this intervention point is already current practice on some services. For example, Snapchat sends users support resources when it finds they have posted content related to self-harm, in addition to removing the content.[935] In response to our 2022 CFE, Meta told us that when an account is reported for posting suicide or self-harm content they may connect the user to organisations that offer help so that they can receive support.[936]

21.228 This evidence suggests that signposting at this intervention point can be an effective way to mitigate the impact of harm to children caused by suicide, self-harm, eating disorder and bullying content. It may also lead to fewer children posting and re-posting such content, thereby reducing the volume of it present on a service and the risk posed to other children from encountering it.

Explanation of the measure in practice

21.229 In light of the evidence mentioned above at paragraph 21.207 that children find immediate signposting to support particularly valuable, we propose to recommend that providers should signpost children to support as quickly as possible when they become aware of children posting or re-posting suicide, self-harm, eating disorder or bullying content (i.e. as quickly as possible following the kind of content being identified). In practice, this might mean signposting children to support some time after the content was originally posted or re-posted, depending on how long it takes providers to detect the content and identify it.

21.230 Providers can be split into two groups: those that already have measures in place that enable them to become aware of when a user posts or re-posts particular kinds of content and those who do not. We discuss each group in turn below.

Providers who already have measures that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder or bullying content

21.231 Many providers already have methods in place that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder or bullying content as part of their content moderation systems. For example, content might be judged to be one of those kinds by a human moderator or flagged as likely to be one of those kinds by an automated content moderation tool.

---

[933] Ofcom, 2024. Experiences of children encountering online content promoting eating disorders, self-harm and suicide.
[934] Ofcom, 2024. Experiences of children encountering online content promoting eating disorders, self-harm and suicide.
[935] Ofcom, 2022. Ofcom's first year of video-sharing platform regulation.
[936] Meta response to 2022 Illegal Harms Call for Evidence.

21.232 As noted above, no content moderation process will be entirely accurate. However, in accordance with Measures CM3 and CM5 in Section 16, providers of large services and services that are multi-risk for content harmful to children should establish performance targets for the accuracy of their content moderation systems and ensure they are well-resourced so as to meet those targets. This would help to mitigate the risk of children being signposted incorrectly.

Providers who do not already have measures that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder or bullying content

21.233 We recognise that not all providers currently have methods in place that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder, or bullying content. For example, providers who choose to remove all content that violates their terms of service may establish that a piece of content violates their terms of service, without identifying the kind of content. See Section 16, Content moderation for U2U services, for our reasons for not recommending use of specific measures to identify these kinds of content at this stage. For the same reasons as set out there, we are not recommending as part of this measure that providers should introduce measures that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder or bullying content.

21.234 In light of this, we are not proposing that providers who do not currently have such measures in place should implement this measure.

Intervention point 3: signpost children when they search for harmful content

Evidence for signposting at this intervention point

21.235 Evidence suggests that children would also benefit from being signposted to support when they search using suicide, self-harm or eating disorder related terms on U2U services. Evidence in Sections 7.2 and 7.3 suggests that some children encounter harmful suicide, self-harm and eating disorder content in this way, something reported more by children with experience of eating disorders, self-harm, suicidal ideation, anxiety or depression.[937] This suggests that signposting children when they search for such content could enable providers to reach those most in need of support.

21.236 We understand that similar features already exist on some services. In response to our 2023 CFE, Pinterest told us that when users search for a blocked term related to suicide or self-harm, they show a suicide and crisis helpline advisory, with no search results. For less sensitive terms, they show the same advisory and include search results, but limit other features such as autocompletion.[938] X similarly told us that when someone searches on their platform for terms associated with suicide or self-harm, the top search result is a notification encouraging them to get support.[939] Snapchat also prompts users towards support if they search for terms (for example, related to suicide or self-harm) that might indicate they are at risk of harm.[940] This suggests that some providers recognise the value of signposting at this stage of the user journey for some kinds of content and already have systems in place to do so.

[937] Section 7.2 Suicide and self-harm content; Section 7.3, Eating disorder content.
[938] Pinterest response to 2023 Protection of Children Call for Evidence.
[939] X (formerly known as Twitter) response to 2023 Protection of Children Call for Evidence.
[940] Ofcom, 2022. Ofcom's first year of video-sharing platform regulation.

21.237 We have not seen evidence that children frequently experience harm from bullying content encountered via search functions on U2U services. We are therefore not proposing to recommend that providers signpost to support for bullying content at this intervention point.

Explanation of the measure in practice

21.238 In light of the evidence mentioned above at paragraph 21.208 that children are likely to benefit more from immediate signposting to support, we propose to recommend that providers should signpost children to support as quickly as possible when they become aware of children searching using suicide, self-harm or eating disorder related search terms (i.e. as quickly as possible following the search terms being identified).

21.239 By suicide, self-harm or eating disorder related search terms we mean search terms that a U2U service provider considers to be general search requests related to suicide, self-harm and eating disorders, and requests seeking specific, practical or instructive information about suicide, self-harm and eating disorders.

21.240 Providers can be split into two groups: those that already have measures that enable them to become aware of when a user searches using suicide, self-harm or eating disorder related search terms and those who do not. We discuss each group in turn below.

Providers who already have measures that enable them to become aware of when a user searches using suicide, self-harm or eating disorder related search terms

21.241 As explained above, some providers already have measures that enable them to become aware of when a user searches using suicide, self-harm or eating disorder related search terms. We understand that this may involve maintaining a list of relevant search terms. We think providers are best placed to determine which search terms or combinations of terms should be included on any such list. However, providers should recognise the changing nature of relevant search terms and the importance of regularly updating any list of relevant terms to reflect this. See Section 8.4, Content promoting suicide (Harms Guidance), Section 8.5, Content promoting self-injury (Harms Guidance), and Section 8.3, Content promoting eating disorders, for further guidance on what providers should consider when developing and maintaining such lists.

Providers who do not already have measures that enable them to become aware of when a user searches using suicide, self-harm, eating disorder related search terms

21.242 We recognise that not all U2U providers currently have measures that enable them to become aware of when a user searches suicide, self-harm or eating disorder related content and we are not proposing that they introduce this. See Section 16, Content moderation for U2U services, for our reasons for not recommending use of specific measures to identify when a user searches using certain search terms at this stage.

21.243 In light of this, we are not proposing that this group of providers should signpost children to support at this intervention point.

Rights assessment

21.244 This proposed measure recommends that services signpost children to appropriate support when they report, post or re-post or search relevant kinds of content harmful to children. We expect that signposting children to support would make it easier for them to access support. As a result, this measure has the potential to help mitigate the risks and impact of

harm posed by harmful content and prevent the most severe outcomes, which is closely aligned with the legitimate aim of the Act in protecting children.

21.245 As with Measure US1, we consider below the potential impacts on users' rights to freedom of expression and association and privacy. As with Measure US1, services may apply this measure to child users only where they use highly effective age assurance to identify child users, or else would need to apply this measure to all users (i.e. including adult users), and we have therefore assessed the potential impact under both scenarios. Where services decide to apply this measure to all users, while not the intended aim of this measure, we believe this may bring benefits to adults by making it easier for them to access support resources if they are affected by suicide, self-harm, eating disorder or bullying content.

Freedom of expression and association

21.246 To the extent that this measure dissuades children (or adults if applied to all users) who are posting or re-posting suicide, self-harm, eating disorder or bullying content from doing so again, or dissuades those who search for such content from going on to encounter it, this is part of the objective of the measure and therefore justified and proportionate to help protect children from these harms.

21.247 We consider that there may be a limited impact on the freedom of expression and association rights of users (including both children and adults where services decide to apply this measure to all users), and those who share beneficial and non-harmful content relating to suicide, self-harm and eating disorders on the service, to the extent that users are also (potentially inadvertently) signposted to support if they report, post or re-post or search for this content and may be dissuaded from posting, re-posting or encountering this beneficial content too. While the presentation of support information may serve as a potential friction in user journeys to that beneficial content, users are not prevented from engaging with the content should they wish to do so. Taking these considerations, and the benefits to children into consideration, we consider that the impact of the proposed measure on the rights to freedom of expression, above and beyond the requirements of the Act, to be limited and proportionate.

Privacy

21.248 We recognise that depending on how service providers decide to implement the proposed measure, it could result in a greater or lesser impact on users' privacy rights under Article 8 of the ECHR as set out in Section 2.

21.249 The proposed measure does not specify that service providers should obtain or retain any specific types of personal data about individual users as part of their implementation of this measure. However, we recognise that the analysis of reported content, posted content or search requests for the purposes of targeting signposted support information to users may involve processing personal data relating to the user who has undertaken this action, although this may well be no more than they ordinarily would process in analysing a report, post or search request under their reporting and content moderation processes in any event. Services which choose to process additional personal data in implementing this measure would need to comply with relevant data protection legislation, including applying appropriate safeguards to protect the rights of both children (who may require special consideration) and adults who would be affected by this measure.

21.250 We therefore consider that the impact of the proposed measure on users' privacy rights to be very limited where services comply with relevant laws, and any interference is necessary and proportionate to secure that providers fulfil their children's safety duties under the Act.

Impacts on services

21.251 Table 21.2 below summarises our assumptions for the direct costs for this measure as a whole. These estimates are for the cost of sourcing the materials for signposting and the implementation of signposting at each intervention point described above.

Table 21.2: Summary of direct cost estimates

Activity	One-off implementation cost	Ongoing annual cost
Sourcing / developing materials	£200 - £25,000	£50 - £6,250
Implementing signposting for children who report	£2,000 - £18,000	£500 - £4,500
Implementing signposting for children who post or re-post	£2,000 - £18,000	£500 - £4,500
Implementing signposting for children who search user-generated content	£2,000 - £18,000	£500 - £4,500

Source: Ofcom analysis

Costs of sourcing or developing appropriate support resources

21.252 Firstly, we have estimated the cost of sourcing or developing appropriate support resources, separately to the cost of introducing a signposting mechanism which we cost below. We have set out these costs for where a service provider sources or develops support resources for all four of suicide, self-harm, eating disorder or bullying content. Services that are only at risk of a subset of these harms would only have to signpost support resources for those harms, and would therefore incur lower costs.

21.253 Providers may choose how they identify or develop the support resources. The low end of our cost estimate would apply where service providers with simple governance structures identify publicly available external support resources produced by expert third parties, which could take around 1 day of professional labour cost to find and sign-off these resources. Costs would be higher when external resources are used but there is a more complex governance process. The high end of our cost estimate reflects providers developing resources internally in consultation with expert third parties, which could take approximately 12 weeks of professional labour cost.

21.254 Using our assumptions on labour costs required for this type of work set out in Annex 12, we would expect the one-off direct costs to be somewhere in the region of £200 to £25,000. We expect that smaller providers are likely to source externally written support resources from expert third parties and incur costs at the lower end of this range. Providers of larger services who decide to develop resources internally in consultation with expert third parties are likely to incur costs towards the higher end of this range. We recognise that costs could be higher than the upper end of this estimate range, potentially substantially so, when services choose to produce resources that are more expensive to create, such as extensive audio-visual materials.

21.255 There would also be some ongoing costs to maintain these resources, such as making sure they are up to date. If the annual maintenance costs were 25% of the implementation cost, then this would be between £50 - £6,250 per annum.

Costs of signposting at intervention point 1: when children report content

21.256 The steps needed to implement this measure would vary based on the design of the service and the complexity of the reporting process and sign-posting functionality. As mentioned in the 'Explanation of the measure in practice' sub-section above, for some services this measure could be implemented in a relatively simple way by adding links to support in an acknowledgement of a report, which we are recommending all services likely to be accessed by children should send following receipt of a complaint (refer to UR3 for further details). More complicated solutions may involve more extensive design of the functionality to signpost at the intervention point; user journey mapping and changes to the UI/UX; and technical development to integrate the signposting with existing reporting workflows, which would include changes to the backend infrastructure. This is likely to require the involvement of various resources, such as graphic designers, web designers, user experience designers, content teams, and developers/engineers, plus Quality Assurance and/or testing teams.

21.257 We have estimated that implementing this measure to signpost children to support when they report specific kinds of content harmful to children could take approximately 1 to 4 weeks of software engineering time, with an equivalent amount of non-engineering time. This could mean one-off direct costs of signposting at this intervention point in the region of £2,000 to £18,000 (in addition to the costs of sourcing signposting materials set out above). Costs will be lower when providers signpost to a range of support resources covering all relevant kinds of content in all cases, and as part of the existing user journey for reporting and complaints. We expect that the costs for smaller services who include signposting to support in an acknowledgement of a report can be close to the lower bound of our estimate. In contrast, costs will be higher when the support is tailored to the specific kind of content a user has reported.

21.258 We would also expect a provider to incur ongoing costs. This would include the cost of maintaining any automated signposting solution and ensuring that any updates to support resources are reflected in the signposting. If the annual maintenance costs were 25% of the implementation cost, then this would be between £500-£4,500 per annum.

Costs of signposting at intervention point 2: when children post or re-post content

21.259 The steps needed to implement this measure would vary based upon the design of the service, the complexity of the signposting functionality of contacting users who have posted or re-posted a relevant kind of content, and the complexity of linking the information on who these users are to this signposting functionality. We expect that the initial implementation may involve designing the functionality to signpost when a service provider is aware of a user having posted or re-posted a relevant kind of content, including user journey mapping and changes to the UI/UX; and technical development to integrate the functionality with existing content moderation workflows, which would include changes to the backend infrastructure. This is likely to require the involvement of various resources, such as graphic designers, web designers, UX designers, content teams, and developers/engineers, plus QA and/or testing teams.

21.260 We have estimated that the direct cost of implementing this measure to signpost children to support when they post or re-post specific kinds of content harmful to children would take approximately 1-4 weeks of software engineering time, with an equivalent amount of non-engineering time in addition to the cost of signposting children who report. Given the time

estimate to implement we would expect the one-off direct costs of signposting at this intervention point to be somewhere in the region of £2,000 to £18,000 (in addition to the costs of sourcing signposting materials set out above).

21.261 We have considered the factors that would lead to services having implementation costs towards the lower or higher end of our estimate. We understand that where systems already exist for automating the content moderation process, adaptations may need to be made to enable details of the posting/re-posting account(s) to be captured when the content is reviewed, incorporating this process into an automated content moderation workflow. For less automated systems, which may be more likely on smaller services, there may be additional processing time for content moderators to capture information on all accounts that have posted the relevant content, creating a higher ongoing cost.

21.262 Where providers do not already have the functionality to contact users who post and reshare harmful content, which may be more likely on smaller services, the addition of this is likely to lead to costs towards the higher end of the range.

21.263 Overall, costs are not necessarily expected to scale with the size of the service and could be high even for some smaller services. We also consider that, although we provide the same indicative time and cost ranges for intervention point 2 as other intervention points, there may be added complexity at this intervention point which increases the likelihood of costs reaching or exceeding the upper bound of our estimates. This includes designing a new user journey to identify the relevant points of intervention, and incorporating the provision of information into relevant workflows, including where users may be contacted some time after they posted or shared the content. Therefore, at this time we have greater uncertainty as to the range of costs associated with this measure.

21.264 In addition to the implementation costs, we would expect a provider to incur ongoing costs. This would include the incremental cost of any signposting solution and ensuring that if support resources are updated these are updated in the signpost. If the annual maintenance costs were 25% of the implementation cost, then this would be between £500-£4,500 per annum. As mentioned, there could be higher ongoing costs for services that do not have automated methods to record information on accounts that have posted and re-posted.

21.265 As set out in the 'Explanation of the measure in practice' sub-section above, we are not recommending as part of this measure that providers should introduce specific systems or processes that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder or bullying content, and so this measure does not entail any additional costs for such activities.

Costs of signposting at intervention point 3: when children search for harmful content

21.266 The steps needed to implement this measure would vary based upon the design of the service and the complexity of the signposting functionality. We expect that the initial implementation may involve designing the signposting functionality, including user journey mapping and changes to the user interface or experience; and technical development to integrate with existing user generated content searching and content moderation workflows which would include changes to the backend infrastructure. This is likely to require the involvement of various resources, such as graphic designers, web designers, front-end designers, content teams, and developers/engineers, plus QA and/or testing teams.

21.267 We estimate that the direct cost of implementing this measure to signpost children to support when they search for certain kinds of user generated content could take approximately 1-4 weeks of software engineering time, with an equivalent amount of non-engineering time in addition to the cost of signposting children who report. We would expect the one-off direct costs for signposting at this intervention point to be somewhere in the region of £2,000 to £18,000 (in addition to the costs of sourcing signposting materials set out above).

21.268 How this might be best achieved will depend on the design of the service. We have given providers flexibility in how they choose to present these support resources, for example providers may choose to implement a banner at the top of the search results or create a pop-up. Providers should design these systems so the support information automatically appears when they become aware of a child searching using suicide, self-harm or eating disorder related search terms. Providers may incur costs at the lower end of the range if they signpost with all available support resources, while providers are likely to incur costs at the high end of this range if they tailor the support resources to the specific search terms that a child uses. We think the flexibility allowed means that smaller services can implement this measure at the lower end of our estimated range.

21.269 In addition to the implementation costs we would expect a provider to incur ongoing costs. This would include the incremental cost of maintaining the system, which may involve quality assurance to ensure that the signpost continues to appear correctly when children search using suicide, self-harm or eating disorder related search terms. If the annual maintenance costs were 25% of the implementation cost, then this would be between £500-£4,500 per annum.

21.270 As noted in the 'Explanation of the measure in practice' sub-section above, we are not recommending as part of this measure that providers should introduce measures that enable them to become aware of when a user searches using terms related to suicide, self-harm or eating disorder content, and so this measure does not entail any additional costs for such activities.

Potential indirect costs

21.271 We recognise that this measure could also result in additional friction for users in terms of additional time and effort that could indirectly be a cost to services. Depending on how it is implemented, on some services users may have to choose either to follow or to ignore the signpost each time they are at a relevant intervention point. This could alter the flow of the user experience, potentially reducing user engagement to some degree and indirectly impacting service revenue.[941]

21.272 However, our measure allows services flexibility to decide how to signpost in a way that is appropriate for their service, which somewhat mitigates this risk. In addition, at the relevant intervention points it is to some degree necessary to interrupt the user experience for the measure to achieve the intended effect. We consider that the introduction of friction where users have encountered harmful content for the purpose of mitigating the impact of this content is an appropriate impact of the measure.

[941] Section 7.12, Business models and commercial profiles, sets out the relationship between engagement and revenue for U2U services.

Which providers we propose should implement this measure

21.273 We expect the benefits of this measure to be material, by intervening at key points of the user journey and reducing harm that may occur where children encounter suicide, self-harm, eating disorder or bullying content.

21.274 There would be costs to providers of implementing this measure, though these would be somewhat mitigated by the flexibility of the measure, which would allow providers to tailor their solutions to their users and platforms. Intervention points 2 and 3 only apply to services that become aware of users posting/re-posting or searching user generated content for the related harms, limiting their scope. In addition, to limit costs while maximising effectiveness, services can signpost to support provided by third parties and not incur costs of developing their own resources.

21.275 The costs and relevant harms vary depending on the intervention point, we therefore consider which providers should signpost at each intervention point in turn separately below.

Intervention point 1: when children report content

21.276 Evidence discussed suggests that signposting at this intervention point soon after children may have encountered harmful content online and been motivated to report it can mitigate the impact of harm posed to children by encountering suicide, self-harm, eating disorder and bullying content.

21.277 We estimate that the costs of signposting at this intervention point are likely to be limited for smaller services due to the flexibility we provide in how services implement this. We estimate the costs would be higher for providers who target the resources shown in each case depending on the specific kinds of content being reported, rather than for those who signpost all children to a range of support for different topics. Given the benefits of signposting children to support, we believe the measure is proportionate for all services with relevant risks, regardless of size.

21.278 As all providers are required to operate reporting processes for content harmful to children, our provisional conclusion is to recommend this measure to all U2U services likely to be accessed by children that are medium or high risk for one or more of suicide, self-harm, eating disorder or bullying content. We propose to recommend these services signpost children to appropriate support when they report content of a relevant kind, for which the service has medium or high risk.[942]

21.279 For the purpose of this measure, relevant kinds of content are: suicide content, self-harm content, eating disorder content, and bullying content.

Intervention point 2: when children post or re-post content

21.280 Evidence discussed suggests that signposting at this intervention point can mitigate the impact of harm posed to children by posting or re-posting suicide, self-harm, eating disorder and bullying content, bringing significant benefits. This has the potential to reach some of those who would benefit most from accessing authoritative sources of support, and may

[942] For example, if a provider has identified that their service is high risk for suicide and bullying content, but not self-harm or eating disorder content, the provider should signpost children to appropriate support when they report suicide or bullying content.

also lead to fewer children posting and re-posting such content, reducing the risk posed to other children from encountering it.

21.281 However, at this stage we have significant uncertainty as to the total cost of this measure, which we anticipate could be more complex and thus costly than the measures proposed at intervention points 1 and 3. In addition, it is not clear at this stage whether smaller services could implement this measure in cost-effective ways. Our analysis suggests that costs for this intervention point 2 do not necessarily scale with the size of the service, and may in fact be greater for smaller services if they have to develop an appropriate capability to contact users who have posted or re-posted content after it has been moderated. On balance, we consider the measure proportionate for providers of large services with the relevant risks and functionalities, but we are not proposing to recommend this measure for smaller services at present.

21.282 We do not propose to recommend as part of this measure that providers should introduce measures that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder or bullying content if they do not already have them. Therefore, signposting at this intervention point only applies where services have such methods. It would also only be relevant for providers of services that enable users to post and re-post content.

21.283 Our provisional conclusion is to recommend this measure to large U2U services likely to be accessed by children that are medium or high risk of one or more of suicide, self-harm, eating disorder or bullying content. We propose to recommend these services signpost children to appropriate support when they post or re-post a relevant kind of content which the service is at risk of, if:

a) They offer users the ability to post or re-post content.
b) They already have methods that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder or bullying content; and
c) They offer users the ability to post or re-post content.

21.284 For the purpose of this measure, relevant kinds of content are: suicide content, self-harm content, eating disorder content, and bullying content.

Intervention point 3: when children search for harmful content

21.285 We have set out evidence that some children encounter harmful suicide, self-harm and eating disorder content by searching for user-generated content, and that these children may be most in need of support. Evidence suggests that signposting at this intervention point can mitigate the impact of harm posed to children by this content. The incremental benefits of this measure are expected to be material, as it addresses a pathway to harm that other measures do not directly address.

21.286 We have not seen evidence that children encounter bullying content via search functions on U2U services. Therefore, while bullying content is considered relevant at intervention points 1 and 2, we do not consider it relevant at intervention point 3.

21.287 The costs of signposting at this intervention point are likely to depend on the complexity of the service, with smaller services that are typically less complex incurring costs at the lower end of our estimated range. Costs may also be higher where providers tailor support to the kind of content children are searching for, and lower where providers signpost all children who search using certain search terms to a range of support for different topics. Services have flexibility in determining how they provide this support following a user's search.

21.288 This intervention point only applies to providers of U2U services that enable users to search for content and can identify when a user searches using terms related to suicide, self-harm or eating disorder related search terms. We believe that services who have this functionality are also likely to have the resources and capability to implement our signposting measure at intervention point 3. This is particularly the case as services have flexibility in determining how they provide this support following a user's search, allowing smaller services to implement this measure incurring costs only at the lower end of our estimates.

21.289 Our provisional conclusion is to recommend this measure to all U2U services likely to be accessed by children that are medium or high risk of one or more of suicide, self-harm, or eating disorder content. We propose to recommend these services signpost children to appropriate support when they search for harmful content using search terms relating to a relevant kind of content which the service is at risk of, if:

a) They already have measures that enable them to identify when a user searches using suicide, self-harm, or eating disorder related search terms; and
b) They offer users the ability to search for user generated content.

21.290 For the purpose of this measure, relevant kinds of content are: suicide content, self-harm content, and eating disorder content.

Other options considered

21.291 We do not consider that at this stage we have sufficient evidence to recommend that providers should signpost children to support when they block content, as was suggested by participants in our research into children's experiences of self-harm, suicide and eating disorder content.[943] Under Measure US4 above, children who restrict their access to content, including through blocking, would be provided with information about other actions they can take, including reporting. If certain kinds of content are reported on providers implementing this measure, the user should then be signposted to support. However, we welcome further evidence on signposting children to support when they use blocking tools and if appropriate may consider this option further as part of our future work.

Provisional conclusion

21.292 We consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes on the basis that it would effectively mitigate the impact of children's exposure to suicide, self-harm, eating disorder and bullying content, on the services we propose should be implementing the measure. For the draft legal text of this measure please see PCU E3 in Annex A7.

[943] Ofcom, 2024. Online Content: Qualitative Research, Experiences of children encountering online content promoting eating disorders, self-harm and suicide.

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

Explanation of the measure

21.293 The Act requires U2U and search service providers to employ user support measures, where proportionate, for the purposes of compliance with the children's safety duties.[944] This measure is intended to ensure that children can benefit from the protection of a service's user-operated safety features.

21.294 In delivering this measure, we would expect to see service providers develop age-appropriate user support materials for children, including explanations for the adults who care for them, ensuring that children can understand the user support tools and reporting and complaints functions on the service, and how to use these to mitigate the risk and impact of encountering harmful content.

21.295 At a minimum, and where services offer any of the following functionalities and processes, the age-appropriate user support materials should explain:

- The option for children to accept or decline invitations to join groups;
- The option for children to block or mute other user accounts;
- The option for children to disable comments on their own posts;
- The process to report harmful content encountered on a service to the service provider;
- The process to submit complaints to a service provider.

21.296 Service providers may consider that there are other tools children can use to support their safety on the service which might also be explained within age-appropriate user support materials.

21.297 To support children's understanding, these materials should be presented in child friendly formats and include visuals, audio-visual elements or interactivity, as well as explanations for parents and carers. They should be presented in ways and at times that promote engagement with the materials.

21.298 For the purposes of this measure, 'age-appropriate user support materials' refer to materials that are specifically designed to be accessible and understandable to all children permitted to use a service, and to the adults who care for them.

21.299 As a baseline, service providers should ensure that the user support materials they produce are both comprehensible to, and emotionally suitable for, the youngest age range permitted to use their service. To ensure that younger children are not exposed to harm-related information more suited to older children, service providers should avoid giving details of harmful content within their user support materials.

[944] Sections 12(8)(g) and 29(4)(e) of the Online Safety Act 2023. The Children's Safety Duties in question are laid out in sections 12(2), 12(3), 29(2), and 29(3) of the Act.

21.300 However, we recognise that no one resource can be targeted to all age groups of children[945] and encourage service providers to consider creating different versions of the user support materials for different age groups of children, allowing children to navigate to the version that suits them best.[946] This consideration might be particularly relevant to large services, or those who permit use by children from a wide range of age groups.[947]

21.301 We are currently only recommending this measure to explain safety features that can protect children from a range of legal content that is harmful to children. More evidence is needed to determine whether age-appropriate user support materials explaining how specific harms from illegal content are addressed would be effective, and we are not therefore proposing to add an equivalent measure to our draft Illegal Content Codes at this time. We welcome evidence and feedback on this.

Effectiveness at addressing risks to children

21.302 This measure aims to increase the effectiveness of a service's user support tools and reporting and complaints processes by ensuring that children can fully understand these provisions and how to use them. At a minimum, this should ensure that where services offer the following functionalities and processes, children have the knowledge and confidence to:

- Accept or decline invitations to join groups;
- Block or mute other user accounts;
- Disable comments on their own posts;
- Report harmful content encountered on a service to the service provider;
- Submit complaints to a service provider; and
- Report potentially harmful predictive search suggestions (for search services).

21.303 This measure will help mitigate and manage the risks and impact of harm to children, by supporting them to access the full protections afforded by these tools and processes in tackling content that is harmful to children, to the extent that these are relevant to and provided by a service.

---

[945] This has been noted in sources including the following: IEEE, 2021. IEEE standard for an age appropriate digital services framework based on the 5Rights principles for children. [accessed 16 April 2024]. Subsequent references are to this document throughout.; 5Rights, 2021. Tick to agree: Age appropriate presentation of published terms. [accessed 16 April 2024]. Subsequent references are to this document throughout.; Save the Children, 2022. How to write a child friendly document. [accessed 16 April 2024]. Subsequent references are to this document throughout.; ICO, 2020. Age appropriate design: a code of practice for online services. [accessed 16 April 2024]. Subsequent references are to this document throughout.; See also: Livingstone, S., Stoilova, M. & Nandagiri, R., 2019. Children's data and privacy online. Growing up in a digital age: an evidence review. [accessed 16 April 2024]. Subsequent references are to this document throughout.; Stoilova, M., Nandagiri, R. & Livingstone, S., 2021. Children's understanding of personal data and privacy online – a systematic evidence mapping, *Information, Communication & Society*, 24 (4). [accessed 16 April 2024].; 5Rights response to 2023 Protection of Children Call for Evidence.

[946] ICO Age appropriate design code, 2020.

[947] We provide guidance in our draft Children's Register of Risks Section 7, and draft Children's Risk Profiles Section 12 about age groups and what service providers should consider when assessing the risk of harm to children in different age groups.

21.304 Providing age-appropriate user support materials for children enables their online safety. Informed children are better able to take appropriate action,[948] for example, when something goes wrong online. This is particularly important if children are repeatedly exposed to online harms and for children who might not have access to engaged adults who can help them to stay safe online.

21.305 Existing guidance,[949] and respondents to our 2023 Protection of Children Call for Evidence (our 2023 CFE),[950] encourage the provision of explanatory materials for children to ensure they can understand information that is pertinent to their online safety.

21.306 Many services already offer guidance and support materials for children and the adults who care for them.[951] For example, in their response to our 2023 CFE, Google note that they provide detailed user-friendly information in their Help Centre about how to make complaints, allowing child users to guide themselves through the reporting process.[952]

21.307 Amazon provide a "Children's Privacy Notice", which is a 90 second cartoon targeted at under 13s.[953] In their Safety Centre, TikTok use pictures and videos alongside text to explain specific aspects of their service to different audiences, including "Privacy Highlights for Teens"[954] and a "Guardian's Guide" for parents.[955] Meta's Safety Centre includes a searchable resource library956 with a specific "Youth" filter, returning partnered and third-party resources on issues relevant to young social media users. Lego have developed a free online game called "Safety Dash", intended to help children and the adults who care for them explore online safety techniques in a gamified setting.[957]

21.308 This measure recommends that service providers meet key characteristics to effectively support children's understanding and engagement when designing their age-appropriate user support materials. Our analysis suggests several characteristics are important in

---

[948] Ofcom, 2022. Serious game pilot: Trialling a serious game as an approach to making children safer online. Subsequent references are to this research throughout. Note: All participants (n=629) were aged between 13 and 17.

[949] "When text requires reading ability more advanced than the lower secondary education level after removal of proper names and titles, supplemental content, or a version that does not require reading ability more advanced than the lower secondary education level, is available." Source: Web Accessibility Initiative, 2023. Web Content Accessibility Guidelines (WCAG) 2.1 W3C Recommendation 21 September 2023 [accessed 16 April 2024]. See also Ofcom, 2021. Video-sharing platform guidance: guidance for providers on measures to protect users from harmful material.; Carnegie UK, 2023. Model code: A reference model for regulatory or self regulatory approaches to harm reduction on social media. [accessed 16 April 2024]. Subsequent references are to this document throughout.; OECD, 2021. Recommendation of the Council on children in the digital environment. [accessed 16 April 2024]; IEEE standard, 2021; ICO Age appropriate design code, 2020.[950] Refuge response to 2023 Protection of Children Call for Evidence; Carnegie UK response to 2023 Protection of Children Call for Evidence; Molly Rose Foundation response to 2023 Protection of Children Call for Evidence.

[950] Refuge response to 2023 Protection of Children Call for Evidence; Carnegie UK response to 2023 Protection of Children Call for Evidence; Molly Rose Foundation response to 2023 Protection of Children Call for Evidence.

[951] Patreon response to 2023 Protection of Children Call for Evidence; Twitter (now X) response to 2023 Protection of Children Call for Evidence; Amazon response to 2023 Protection of Children Call for Evidence; Meta response to 2023 Protection of Children Call for Evidence.

[952] Google response to 2023 Protection of Children Call for Evidence.

[953] Amazon, Children's Privacy Notice. [accessed 17 April 2024].

[954] TikTok, Privacy Highlights for Teens. [accessed 17 April 2024].

[955] TikTok, Guardian's Guide. [accessed 17 April 2024].

[956] Meta, Safety Centre Resource Library. [accessed 17 April 2024].

[957] Lego, Play Safe Online. [accessed 17 April 2024].

ensuring that children can understand, and are likely to engage with, such materials online. We explore these characteristics in more detail below.

Understanding age-appropriate user support materials

Providing materials in child-friendly formats

21.309 It is important that children can independently access and understand materials explaining the tools available to help them feel safer on a service, particularly where they do not have, or do not want, adult support.[958] 5Rights found that just nine out of 123 websites likely to be accessed by children had privacy policies targeted at children, although several had policies aimed at the parents of under-13s.[959]

21.310 Respondents to our 2023 CFE,[960] as well as relevant guidance,961 repeatedly recommend that services provide audio-visual and even interactive resources for children to ensure they can understand otherwise text-based information.

21.311 The ICO's Children's code advises that effective formats for presenting information to children in a child-friendly way can range from diagrams, cartoons and graphics, through video and audio content to gamified or interactive content.[962] When consulted by 5Rights, a workshop group of 12–16-year-olds expressed a preference for written information to be presented in easier formats, such as animations, audio, video or graphics.[963] Our own research has found that children prefer to see information about a platform in the form of short videos or images, with detailed text their least preferred format.[964]

21.312 Importantly, more engaging formats (e.g. audio-visual and interactive materials) have been found to improve comprehension, and subsequent online safety behaviours, among children when compared to written information. In a pilot study, we found that among 13–17-year-olds, an interactive and visually stimulating 'serious game' (an online quiz-style game aimed at educating children) improved knowledge and understanding of social media etiquette more than text-based guidance. Participants were also more likely to implement positive social media etiquette in the two weeks after playing the serious game.[965] Other recent research from Ofcom found that 13-17-year-olds who were exposed to age-appropriate user support materials using a hybrid of images and short text had significantly better comprehension of available support tools than those who had not been exposed to these materials. [966]

---

[958] 5Rights response to 2023 Protection of Children Call for Evidence.
[959] 5Rights Tick to agree, 2021.
[960] ParentZone response to our 2023 Protection of Children Call for Evidence; Antisemitism Policy Trust response to 2023 Protection of Children Call for Evidence; ICO response to 2023 Protection of Children Call for Evidence; SWGfL response to 2023 Protection of Children Call for Evidence; UKSIC response to 2023 Protection of Children Call for Evidence.
[961] Designing for Children's Rights, 2022. Design Principles: Version 2.0. [accessed 16 April 2024]; Schneble, Favaretto, Elger & Shaw, 2021; Save the Children, 2022.
[962] ICO Age appropriate design code, 2020.
[963] 5Rights Tick to agree, 2021. Note: The young people consulted were between 12 and 16 years of age (group size unknown).
[964] Ofcom, 2024. Ofcom: Engaging with User Support Materials Trial. Subsequent references to this document throughout. Q: How would you like to see information about a platform's rules or how to do things in a help centre?: Short videos (52%), Images (42%), Short articles (32%), Tours or demonstrations (28%), Infographics (24%), Interactive games (21%), Detailed text (19%).
[965] Ofcom serious game pilot, 2022.
[966] Ofcom engaging with user support materials, 2024.

21.313 Our research testing the effects of prompts and nudges to encourage reporting via micro-tutorials in an adult population revealed that when exposed to a micro-tutorial (static, audio-visual video, or interactive), reporting of potentially harmful content significantly improved, showing the benefits of providing users with support materials. [967] However, audio-visual video and interactive micro-tutorials were most effective, showing the positive impact of the format of information. Given the results of the 'serious game' and user support materials research, we might expect similar reactions to interactive micro-tutorials among children.

21.314 Presenting information about relevant user tools in more engaging, child-friendly, formats should therefore be effective in helping children to understand what tools are available and how to use them to stay safe on a service.

Providing parental guidance

21.315 Parents and other adult caregivers can play an important role in protecting children online.[968] In recent findings from their twice-yearly tracking survey, Internet Matters found that 48% of children who had experienced harm online went to their parents/guardian to discuss it, while 9% had a conversation about it with their teacher.[969]

21.316 However, when they reviewed existing knowledge on children's data and privacy online, researchers at the London School of Economics found notable gaps in adults' understanding of online risks to children, advising that media literacy resources and training should be provided for parents, educators and child support workers.[970]

21.317 Respondents to our 2023 CFE,971 as well as existing research and guidance,[972] agree that information for parents/guardians should be made available, with some services already providing such guidance.[973] For example, Lego provide both online and offline resources to help parents/guardians facilitate conversations with their children about staying safe online.[974]

21.318 Providing guidance materials for parents/guardians can help them understand the risks to children online and how these can be managed on a service. These adults are then better placed to explain or explore this with the children in their care, ensuring the children know how to employ user tools to feel safer online.

---

[967] Ofcom, 2023. Boosting users' safety online: Microtutorials
[968] Internet Matters response to 2023 Protection of Children Call for Evidence.
[969] Internet Matters, 2023. Insights into children's digital user: November 2023 tracker survey. [accessed 17 April 2024].
[970] Livingstone, Stoilova & Nandagiri, 2019. Children's data and privacy online: growing up in a digital age: an evidence review. [accessed 26 April 2024].
[971] ParentZone response to 2023 Protection of Children Call for Evidence; Internet Matters response to 2023 Protection of Children Call for Evidence; Centre for Countering Digital Hate response to 2023 Protection of Children Call for Evidence; Executive Office NI response to 2023 Protection of Children Call for Evidence.
[972] Schneble, C.O., Favaretto, M., Elger, B.S. & Shaw, D.M., 2021. Social media terms and conditions and informed consent from children: Ethical analysis, *JMR Pediatrics and Parenting*, 4 (2). [accessed 16 April 2024]. Subsequent references are to this research throughout.; ICO Age appropriate design code, 2020.
[973] Schneble, Favaretto, Elger & Shaw, 2021; Twitter (now X) response to 2023 Protection of Children Call for Evidence; Patreon response to 2023 Protection of Children Call for Evidence; Meta response to 2023 Protection of Children Call for Evidence.
[974] Lego, Raising digitally smart families. [accessed 17 April 2024].

21.319 This may be particularly important for children who are unable to independently access guidance materials without support, perhaps because of their young age, or a disability.

Engaging with age-appropriate user support materials

Presented during sign-up

21.320 Children must be made aware of age-appropriate user support materials in order to engage with them. This should happen as early as possible in the user journey, to ensure that children know what user tools are available to them as soon as they begin using a service, and to increase awareness of available support materials so children can revisit them at a later point in their user journey.

21.321 In recent research,[975] we found that around a third (35%) of 13-17-year-olds clicked through to view age-appropriate user support materials while signing up to a mock platform versus less than 1% who saw a static link to a help centre summarising the terms of service. This suggests that children are likely to engage with age-appropriate user support materials from the earliest stages of service use if they are made aware of the materials in a prominent and engaging way. Further, participants who received a salient prompt to age-appropriate user support materials during sign-up were also four times more likely to recall at a later stage that the materials existed, suggesting early and clear prompting of these materials could benefit children later in their user journey if they were in need of support.

21.322 The research also revealed that almost half of participants thought it most important to understand information about the platform before, or during, sign-up.[976] This suggests children would welcome early awareness of age-appropriate user support materials.

21.323 We recognise that most search services do not require users to sign- before they can engage with the search engine. However, where they do provide users with the option to sign up, we would expect the provider to ensure that the materials are made available during this process.

Can be returned in internal search results

21.324 Around half of the children who participated in our recent research preferred to understand information about a platform as needed after sign-up.[977] In a separate piece of qualitative research, children said they were more likely to use Google or speak to an authority figure such as a parent or group admin about how to report content than to look for it on a platform.[978] This suggests user support materials should be easily findable on a service in order to encourage engagement with them.

21.325 Published guidance on presenting age-appropriate information advises that key terms and definitions should be searchable, allowing children to explore returned results relevant to

[975] Ofcom engaging with user support materials, 2024.
[976] Ofcom engaging with user support materials, 2024. Q: When, if at all, do you think it's most important for users to understand the following? General information about the platform and how to use it: 47% before or during sign-up.
[977] Ofcom engaging with user support materials, 2024. Q: When, if at all, do you think it's most important for users to understand the following? General information about the platform and how to use it: 23% after sign-up but before commenting or posting, 25% on a help centre when needed.
[978] Ofcom, 2024. Children's Attitudes to Reporting Content Online.

their query. This would enable children to quickly access the specific materials they required without having to navigate to, and then through, a service's help centre. [979]

21.326 Engagement with age-appropriate materials would likely be increased if they were returned as relevant internal search results. This would in turn increase understanding of services' features and functionalities among children, allowing them to easily and repeatedly find digestible information about using the service.

21.327 We note that not all search services will have a separate 'internal' function through which users can search for material produced by the provider, such as policies on reporting and complaints procedures, and where they do, these may not be as prominent as the primary search engine of which the search service consists. In these instances, we would therefore expect that search services should provide the relevant age-appropriate information in response to search queries entered using the primary search engine. Where they have a separate internal search function, the material should also be provided in that context.

Accessible to users and non-users

21.328 In line with our recommendation around Terms of service and publicly available statements Section 19, being able to access key information about a service prior to sign-up is important to ensuring its suitability for children.[980] If children and the adults who care for them can see age-appropriate information about a service's safety tools before becoming a user, they can make an informed decision about whether to sign up for a service.

21.329 The materials should be accessible without needing a user account. We note that this is generally the practice of search services, as those services do not typically require users to sign up in order to use the service and policies and guidelines tend to be publicly available via URLs. However, for U2U services, this would require service providers to make the materials indexable, allowing them to be returned in external search service results. To make the materials available to non-users, app-only services may choose to present them on a webpage.[981]

Rights assessment

21.330 This proposed measure recommends that providers develop age-appropriate user support materials for children, ensuring that they understand the user support tools and reporting and complaints functions on the service, and how to use these to mitigate the risk and impact of encountering harmful content. The aim of this measure is to ensure that children can benefit from the protection of a service's user-operated safety features. As noted above, the Act requires services to employ user support measures, where proportionate for the purposes of compliance with the children's safety duties.

21.331 We have carefully considered whether this proposed measure would constitute an interference with users' (both children and adults) or services' freedom of expression or association rights, or user's privacy rights. Our provisional conclusion is that it would not.

---

[979] IEEE standard, 2021; 5Rights Tick to agree, 2021; 5Rights response to our 2023 Protection of Children Call for Evidence.
[980] Carnegie UK Model Code, 2023; Carnegie UK response to 2023 Protection of Children Call for Evidence.
[981] Organisations often meet the UK GDPR requirement to make privacy information easily accessible by putting this information on their website. Source: ICO, When should we provide privacy information?. [accessed 17 April 2024]. This suggests that app-only services in scope of GDPR may already have a web presence. We found this to be the case for BeReal and WhatsApp.

This proposed measure only requires services to develop user support materials to aid users' understanding of the safety tools which may feature on a service. The measure itself does not require any steps to be taken with respect of particular kinds of content nor does it require the use of any personal data. We additionally consider that age-appropriate user support materials may have positive impacts on users' - particularly children's – rights to freedom of expression and association, and also their rights to privacy, in that it should also help them to understand the options they have to protect themselves from encountering content or contacts that might be harmful to them, and protect their personal data, as they use the service to express themselves and connect with other users.

Impacts on services

21.332 In-scope service providers that do not currently have age-appropriate user support materials would need to develop them. We have considered the expected cost implications of this below. The detailed assumptions underlying our direct cost estimates are found in Annex 12.

Table 21.3: Summary of direct cost estimates

Activity	One-off implementation cost	Ongoing annual cost
Producing materials - research phase	£1,000 - £26,000	£250 - £6,500
Producing materials - implementation of the measure / design phase	£2,000 - £47,000	£500 - £11,750
Producing materials – deployment phase	£2,000 - £20,000	£500 - £5,000
Engagement with materials	£1,000 - £3,000	£250 - £750
Total cost	**£6,000 - £96,000**	**£2,000 - 24,000**

Source: Ofcom analysis

Costs linked to producing age-appropriate user support materials

21.333 The bulk of the costs associated with this measure would depend on the type and quantity of parental explainers and child-friendly materials that services decide to create. This cost would be determined by how a service chooses to create the materials explaining user support tools and reporting and complaints procedures, the number of user support tools that the service needs to create these materials for, as well as whether the service decides to provide different versions of these materials targeted at different age groups of children.

21.334 Given the variety of type and quantity of materials that services can choose from, we estimate a lower and a higher cost estimate. Each estimate includes both research and implementation costs.

21.335 Services would need to conduct research to ideate the design of the user support materials appropriate for their audience. We estimate that at the lower end, conducting basic research to produce simple materials like pictures or audio for a single user tool (i.e. reporting and complaints procedures, which is the only tool that would apply to any service) would entail five working days from a UX designer or researcher, with costs between £1,000 and £3,000. At the higher end, the research costs could also include interviews or surveys with children (and analysing resulting data) as well as requesting expert advice. At the higher

end, we estimate that the one-off research costs would entail 50 working days from a UX designer or researcher. We estimate these costs to be between £13,000 and £26,000.[982]

21.336 The implementation of this measure would first require a design phase and as mentioned, the activities that services would need to undertake in this phase would depend on the type(s), quantity, and quality of materials they decide to produce. At a minimum, this phase could include creating storyboards, diagrams, cartoons or comic strips, recording audio, etc. At the lower end, we expect one UX / UI designer and one content creator and each would spend five working days each in this phase. We estimate these costs to be between £2,000 and £4,000. This phase could also include scriptwriting, animations and graphics, planning sound design, recording audios, preparing gamified or other interactive material, conducting usability testing etc. At the higher end, we expect UX / UI designers, content creators, and graphic and multimedia designers to spend 40 working days each in this phase. We estimate these costs to be between £24,000 and £47,000.

21.337 The design phase is likely to be followed by the deployment phase, which would require back-end development to store the materials created and front-end development to display the materials created. This phase would also require user testing and debugging. The costs in this phase entail the time of a software developer or engineer as well as cost of oversight from areas like project management, trust and safety, legal, and policy. At the lower end, we estimate three working days of a software engineer's time in this phase with an equal amount of non-software engineering time (e.g. project management, legal, trust and safety, policy), with costs between £2,000 and £3,000. At the higher end, we estimate 20 working days of a software engineer's time in this phase with an equal amount of non-software engineering time (e.g. project management, legal, trust and safety, policy), with costs between £10,000 and £20,000.

21.338 In-scope services would also incur maintenance costs to keep the materials up to date with any changes to their user support tools or reporting and complaints procedures and to reflect any newly adopted tools. We assume annual maintenance costs to be 25% of the implementation costs and estimate these to be between £2,000 and £3,000 per year. At the higher end, we estimate these costs to be between £12,000 and £24,000 per year. Overall, we anticipate that the costs associated with implementing this measure are likely to represent a higher share of revenue for smaller services. However, the cost burden on smaller services would be mitigated to some extent by the flexibility allowed by the measure over the types of materials they wish to produce. We also expect these costs to scale with the risks present on a service as riskier services might already have or might need to adopt more user support tools and therefore incur more costs in developing user support materials explaining how to use these tools.

One-off costs linked to ensuring enable engagement with age-appropriate user support materials

21.339 As part of the proposed measure, services would also need to promote engagement with the user support materials (by presenting the materials during sign-up, making them searchable on the service, and making them accessible to both users and non-users). We estimate around five working days of a software engineer's time to ensure this and estimate the one-off costs of this to be between £1,000 and £3,000.

[982] We provide a range based on different salary ranges as set out in Annex 12

Potential indirect costs

21.340 We recognise that presenting materials during sign-up as part of this measure could result in costs to users in terms of additional time and effort. This could alter the flow of the user experience, potentially reducing user engagement to some degree and indirectly impacting service revenue.[983] However, our measure allows services flexibility to decide how to present the materials in a way that is appropriate for their service, which somewhat mitigates this risk. On balance we consider it unlikely that this measure would materially discourage a high proportion of users from using the service, considering that many internet users are accustomed to dealing with different kinds of prompts or notification. We consider any indirect costs to be acceptable considering the benefits of ensuring that children and adults who care for them are made aware of the availability of these materials.

Overall cost

21.341 Overall, we expect the one-off costs linked to producing materials and ensuring engagement with them to fall between £6,000 and £13,000 at the lower end. This is likely to be the range of costs for services which choose to produce basic or simpler user support materials and have fewer risks since they would need to adopt fewer user support tools and therefore produce fewer materials. We recognise that the benefits of this measure would increase as a service chooses to produce higher cost and higher quality materials.

21.342 We estimate the costs to be £48,000 and £96,000 at the higher end. This is likely to be the range of costs for services which might want to produce more advanced types and varieties of user support materials and riskier services which might need to adopt more user support tools.

21.343 Annual maintenance costs are expected to fall between £1,000 and £3,000 at the lower end and between £12,000 and £24,000 at the higher end.

Which providers we propose should implement this measure

21.344 We have provisionally concluded to recommend this measure to providers of all U2U and search services likely to be accessed by children that are multi-risk for content harmful to children.

21.345 We believe this measure supports the effectiveness of a service's reporting and complaints processes and other user support tools that a service implements to mitigate risks and improve safety outcomes for children. It does so by ensuring that children on risky services and the adults who care for them clearly understand how children can use these tools. This comprehension is key in ensuring that children feel empowered to make use of the tools at their disposal and can safely navigate the risks that might be present on services, ultimately making them safer online.

21.346 Given this benefit, we consider it proportionate to recommend this measure to services that are multi-risk for content harmful to children. The impact on children's safety from this measure is expected to be material for multi-risk services, as such services are more likely to have a range of relevant user tools to cover in the materials, and children are more likely to benefit from understanding better how they can manage their risk of exposure to these different harms. We believe that the flexibility we allow in how services can practically

[983] Section 7.12, Business models and commercial profiles, sets out the relationship between engagement and revenue for U2U services.

implement this measure would ensure that it is appropriate to their circumstances, capabilities and financial resources. Since we expect riskier services to develop more user support tools, the benefits of this measure increase with the risks present on a service. This makes the costs incurred proportionate to the anticipated benefits of explaining to children and the adults who care for them the tools that they can use to be safe from the risks associated with the service.

21.347 We recognise that some services that are multi-risk for content harmful to children may have more limited user tools and this measure may only be relevant to explaining a more narrow subset of functionalities, such as the service's reporting and complaints processes.[984] While benefits will be lower compared to services with multiple user support tools, costs would also be lower because costs increase with the number of user support tools covered. We consider that benefits would still be significant to justify costs as improving the effectiveness of the complaints process has the potential to reduce risks in relation to multiple types of content harmful to children.

21.348 At this stage we do not consider it proportionate to recommend this measure for services that are not multi-risk for content harmful to children. For the same reasons set out above, we expect that benefits would be limited for these services. While there are potentially some benefits for single-risk services and the costs of this measure in isolation could be manageable for some of them, we have considered the combined implications of this measure on top of others. As set out in our combined impact assessment Section 23, we consider that the overall cost burden on some single-risk services may negatively affect users and people in the UK, so we have prioritised other measures for them where the benefits are more material.

Other options considered

21.349 We considered taking a prescriptive approach to this measure, recommending that service providers create specific kinds of age-appropriate user support materials tailored to specific age-groups of children. However, given the diversity and complexity of the services implementing this measure, including their user bases, the design of their service, and their available resource, we do not consider that a prescriptive approach offers enough flexibility to ensure that the most suitable and effective age-appropriate user support materials would be provided across every service implementing the measure.

21.350 Instead, we recommend that service providers achieve outcomes in line with the five characteristics supporting children's understanding of, and engagement with, age-appropriate user support materials as set out above. We are confident that this would make our broad expectations clear to service providers, while allowing them more flexibility in the steps that could be taken to create suitable and effective age-appropriate user support materials.

[984] For example, the reporting and complaints processes are the only elements of this measure relevant to some search services (including those without a predictive search functionality) as the other user support tools covered by this measure are not relevant to them. Therefore, the benefits of this measure would only be in relation to explaining the reporting and complaints process (unless they choose to cover other user support tools relevant to their service). Similarly, there may still be some user-to-user services that are multi-risk for content harmful to children but are not in scope of any of the other user support tools covered by this measure.

Provisional conclusion

21.351 We consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes due to the benefits it would provide in protecting children from harmful content as explained above. For the draft legal text of this measure please see PCU E1 in Annex A7 and PCS E1 in Annex A8.

22. Search features, functionalities and user support

Search services can act as a pathway to harm. Search features and functionalities, such as predictive search, that have been designed to enhance the user search experience, can, in some circumstances, increase the risk of children being exposed to, and encountering, PPC and PC.[985]

This section details our recommendations for the design of search services to help them meet their safety duties in sections 29(2) and 29(3) of the Online Safety Act 2023 ("the Act"). We believe these measures will have significant immediate impact on minimising children's risk of exposure to PPC and PC.

We propose to adopt an approach consistent with that outlined in our previous 2023 Illegal Harms Consultation. The measures to protect children that we propose in this chapter, however, should be considered separately, and in addition, to those outlined in the Illegal Harms consultation. That is because there are differences in the duties underlying these measures that are unique to protecting children from harm.

Our proposals

#	Proposed measure	Who should implement this[986]
SD1	**Offer users a means to easily report predictive search suggestions relating to PPC and PC**	All large general search services
SD2	**Provide crisis prevention information in response to known PPC-related search requests regarding suicide, self-harm and eating disorders**	All large general search services

Consultation questions

54. Do you agree with our proposals? Please provide underlying arguments and evidence to support your views.
55. Do you have additional evidence relating to children's use of search services and the impact of search functionalities on children's behaviour?
56. Are there additional steps that you take to protect children from harms as set out in the Act? If so, how effective are they?

As referenced in the Overview of Codes, Section 13 and Section 17, Search moderation, the use of GenAI to facilitate search is an emerging development and there is currently limited evidence on how the use of GenAI in search services may affect the implementation of the safety measures as set out in this section. We welcome further evidence from stakeholders on the following questions:

[985] See Volume 3, Section 7.10, Risk of harm to children on search services.
[986] These proposed measures relate to providers of services likely to be access by children.

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.

Search features and functionalities

22.1 In Volume 3, Section 7.10, Risk of harm to children on search services, we explain that search services have designed and implemented features and functionalities to enhance the user search experience. Evidence indicates that some search functions, such as predictive search functionalities can lead users, including children, to encounter harmful content.[987]

Definition box 1: Search related terminology	
General search services	Enable users to search the contents of the web by inputting search requests on any topic and returning relevant results.
General search services which rely on their own indexing	Some general search services rely solely on their own indexing, using crawlers ('crawling') to find content across the web, building an index of URLs ('indexing') and using algorithms to rank the content based on relevance of the search request ('ranking'). General search services are also integrating GenAI to support or perform search functions, for example, by integrating a large language model to provide a conversational summary of that search results.
Vertical search services	Also known as 'speciality search engines,' they enable users to search for specific topics, products or services offered by third party providers. They operate differently to general search services; rather than crawling the web and indexing webpages, it presents users with results from selected websites. Vertical search services may have a contract and API with selected websites or equivalent technical means.
Predictive search	Is an algorithmic feature embedded in the search field through which a search service anticipates or predicts a user's search request based on a variety of ranking factors and provides a list of suggested search requests (referred to as 'predictive search suggestions').
Crisis Prevention information	Refers to information provided by a search service in search results that typically contains the contact details of helplines and/or hotlines and links to trustworthy and supportive information provided freely by a reputable and reliable organisation.

[987] In Section 7.10, Risk of harm to children on search services, several sources of evidence are referenced demonstrating the role of autocomplete in aiding searches for types of potentially illegal content, and it is reasonable to assume the functionality works similarly for searches of content of all types.

Risks associated with search features and functionalities

22.2 Volume 3, Section 7.10, Risk of harm to children on search services and Section 17, Search moderation, distinguish between, and detail the risks presented by different search service types, including general search services and vertical search services. General search services can, in particular, act as a gateway to content that is harmful to children and pose a greater risk of harm in comparison to vertical search services.

22.3 Search services are distinct from U2U services in that they facilitate access to a wide range of websites or databases and can, with minimal friction, provide access to large volumes of content that may be harmful to children. This is particularly the case if users are actively or deliberately searching for such content. This includes PPC, such as content that encourages, promotes or provides instructions for suicide, self-harm and eating disorders.[988]

22.4 Evidence indicates that searching for suicide, self-harm and eating disorder content can return large volumes of content and this content often appears high up in search results, including on the very first page. Research from the Network Contagion Research Institute (NCRI) showed that major search services returned large volumes of content if users actively searched for content considered harmful to children, including suicide, self-harm and eating disorder content. The use of coded language such as abbreviations and cryptic words in search requests (i.e., those used by online communities who create and share this harmful content) generated the most results considered to be harmful. The report found that 1 in 5 search results tested returned content that promoted self-injurious behaviour (including content related to eating disorders). These results often appeared in the first five search results.[989] A 2021 study by Borge et al had similar findings: search requests for general suicide related terms and related to specific suicide methods returned a range of "harmful" results within the first 20 search results.[990]

How children use search services

22.5 While there is limited evidence regarding children's search behaviour and experience on search services, Ofcom research has found that some children actively search for content on eating disorders, self-harm, and suicide on social media platforms.[991] A minority of children from Ofcom research also said that they had inadvertently come across violent content via search services due to making a mistake while searching for something else.[992]

---

[988] Of 37,647 search results reviewed resulting from 450 search queries, researchers classified 22% as containing content that clearly promoted self-injurious behaviour (related to eating disorders, suicide or non-suicidal self-injury). Ofcom, 2024. One Click Away: A Study on the Prevalence of Non-Suicidal Self Injury, Suicide, and Eating Disorder Content Accessible by Search Engines

[989] Ofcom, 2024. One Click Away: A Study on the Prevalence of Non-Suicidal Self Injury, Suicide and Eating Disorder Content Accessible by Search Engines.

[990] The study found that 22% of Microsoft Bing URLs, 19% of DuckDuckGo URLs and 7% of Google Search URLs were "harmful, meaning determined by the researchers as encouraging, promoting, or facilitating suicide or containing discussions of suicide. Borge et. al., 2021, How Search Engines Handle Suicide Queries.

[991] Participants in this study included children and young people aged 13-21, those aged 18+ were reflecting back to their experiences as children. Ofcom, 2024. Online Content: Qualitative Research. Experiences of children encountering online content relating to eating disorders, self-harm and suicide.

[992] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children. Qualitative Research Report.

22.6 Children are also aware of codewords for PPC and PC-related harms that is less likely to be detected by U2U service moderators and that which bypass support or signposting restrictions.[993] Young people with lived experiences reported being more familiar with codewords, and described using them to search for content on U2U services which might otherwise be restricted, such as suicide, self-harm and eating disorder content.[994] Though indicative of young people's search experience on U2U services, this research highlights children's search behaviour and intent in searching for content specifically related to suicide, self-harm and eating disorders. We recognise there is some evidence that children search for this content on search services.

22.7 Research has found that young people specifically make internet searches for suicide methods and ideas that are likely to return harmful content. A study that investigated 145 cases of suicide in young people, including children, under 20 years of age, found that Internet use related to suicide (i.e., internet searches for suicide methods, suicidal ideas posted on social media, or online bullying) was recorded in 30 (23%) deaths. Of the 16 individuals who had searched the internet for information about suicide methods, five died by a method they were known to have searched.[995]

22.8 We have limited evidence indicating the existing prevalence of PC on general search services. However, our Illegal Harms Consultation included evidence on the availability of certain categories of illegal content via search services that overlap with PC-harms, such as hate material and racist content. Section 7.10, Risk of harm to children on search services and Section 17, Search moderation, explains our understanding that general search services operate by indexing most of the webpages across the 'clear web.' In practice, this means that any content, including PC, which has been indexed can be presented in search results if enabled by the ranking system. As such, we assess that children can encounter, and be exposed to, PC on general search services.

22.9 We propose that our measures should apply to large[996] general search services and do not currently assess that smaller general search services or vertical search services should be in scope. As general search services act as a gateway to the entire content of the internet it is possible that children could use them to access, either inadvertently or deliberately, content that is harmful to children, including PPC and PC. We have excluded vertical search services from the scope of these proposed measures as there is no clear evidence they pose a risk to children encountering PPC and PC.

Risks associated with predictive search functionalities

22.10 Predictive search functionalities can be a helpful and time-saving tool designed to improve the search experience by anticipating search requests based on several factors, including popularity and search history.[997] Predictive search can improve the search experience of all users, particularly vulnerable users, including those with dyslexia or cognitive or motor

[993] Ofcom, 2024. Online Content: Qualitative Research. Experiences of children encountering online content relating to eating disorders, self-harm and suicide.
[994] Ofcom, 2024. Online Content: Qualitative Research. Experiences of children encountering online content relating to eating disorders, self-harm and suicide.
[995] Rodway et al. 2021. Suicide in children and young people in England: a consecutive case series.
[996] See Framework for Codes at Section 13 within this Volume for a definition of a large service.
[997] The Illegal Harms Code of Practice identified that predictions are based on many factors including past user queries, location and trends. Ofcom 2023, Volume 4 How to mitigate the risk of illegal harms - the illegal content Codes of Practice Chapter 22 paragraph 22.8

disabilities.[998] However, evidence indicates that predictive search functionalities can lead users, including children, to encounter illegal content[999] including hate speech, CSAM and fraud-related content.[1000]

22.11 There is a risk that search prediction may lead users to harmful content that they might otherwise not have encountered had the search suggestions not been surfaced.[1001] Samaritans' response to our 2022 Illegal Harms Call for Evidence suggested that predictive search can increase the discoverability of harmful suicide and self-harm content, and it recommended that "autocomplete searches [are] turned off for harmful searches such as those relating to methods of harm and associated equipment."[1002]

22.12 While evidence on the potential risks presented by predictive search does not cover every category of PPC and PC, we provisionally consider that it is reasonable to assume that predictive search could facilitate children encountering the full range of search content that is PPC and most categories of PC by virtue of how they operate.

22.13 We acknowledge that content associated with some PC-harms, such as bullying, is less likely to surface via predictive search functions given evidence of how bullying manifests online. As per Section 7.5, Bullying content, research suggests that bullying content is particularly likely to occur on social media and gaming services. This may be because bullying content is often targeted against a person and functionalities such as direct messaging and commenting can facilitate these interactions.[1003] We believe, however, that where any category of PC exists on the web, such as on social media sites and other online forums, predictive search could facilitate children encountering this content, including bullying content.

---

[998] Kennecke, Ann-Kathrin, Wessel, Daniel and Heine, Moreen, 2022. Dyslexia and Accessibility Guidelines – How to Avoid Barriers to Access in Public Services. [Accessed 10 December 2023]

[999] In Ofcom's Register of Risks for Illegal Content, several sources of evidence are referenced demonstrating the role of autocomplete in aiding searches for types of potentially illegal content, and it is reasonable to assume the functionality works similarly for searches of content of all types. Ofcom 2023, Volume 2: The causes and impacts of online harm. Chapter 6U Paragraph 6U.47

[1000] A study found that Google recommended search suggestions of a violent and offensive nature against members of the LGBTQ+ community, when non-offensive words were typed into its search bar.(Google stopped recommending such phrases a week after these examples were flagged). Loeb, J., 2018. Google is 'promoting hate speech', claims internet law expert, E&T, 22 January. [accessed 7 December]; Ofcom, 2023. Online content for use in the commission of fraud – accessibility via search services . [accessed 22 September 2023];

[1001] The Antisemitism Policy Trust reported that Microsoft Bing directed users to hateful searches with autocomplete suggestions containing antisemitic phrasing. Antisemitism Policy Trust response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.; Ofcom, 2023. Articles and items for use in the commission of fraud – accessibility via search services. [accessed 7 November 2023]; Microsoft, When Are Search Completion Suggestions Problematic? P171:17 [accessed 13 November 2023]; Ofcom 2023, Volume 4 Illegal harms consultation Chapter 22 paragraphs 22.8 – 22.18 ;

[1002] Samaritans response to 2022 Ofcom Call for Evidence: First phase online safety regulation

[1003] See [Section 7.5, Bullying content for evidence on how this harm manifests online.

Interaction with Illegal Harms

22.14 In our Illegal Harms Consultation we proposed the following measures regarding search service design be included in our draft Illegal Content Codes:

a) **Measure 1:** Services that use a predictive search functionality should offer users with a means to easily report predictive search suggestions which they believe can direct users towards priority illegal content.
b) **Measure 2:** Provide crisis prevention information in response to search requests that contain general queries regarding suicide and queries seeking specific, practical or instructive information regarding suicide methods.
c) **Measure 3:** Employ means to detect and provide warnings in response to search requests the wording of which clearly suggests that the user may be seeking to encounter CSAM.

22.15 See Section 22 of our Illegal Harms Consultation for a detailed discussion of the evidence and impacts of those measures.

22.16 We provisionally consider that measures 1 and 2 in the draft Illegal Content Codes are also proportionate for providers of a service likely to be accessed by children when adjusted to cover PPC and PC, as relevant. We provisionally consider that Measure 3 does not translate in the same way to the children's safety duties. We set out below our detailed assessments of the evidence and impact of these measures as they relate to duties for services likely to be accessed by children.

Our proposals to protect children

22.17 The Act requires that search services take steps to minimise the risk of children encountering PPC, PC and NDC.[1004] As part of these safety duties, service providers should take steps, where proportionate, relevant to:

a) the design of functionalities, algorithms, and other features relating to the search engine (section 29(4)(a)),
b) functionalities allowing control (especially by children) of the content that is encountered in search results (section 29(4));
c) content prioritisation (section 29(4)(d)); and
d) user support measures (section 24(4)(e)).

22.18 Our proposals focus on the operation and design of search functionalities and user support measures, and content prioritisation. They are aimed at minimising the risk of children encountering PPC and PC, not NDC.

22.19 As set out Section 7.10, Risk of harm to children on search services, evidence suggests that some search functionalities are particularly effective way for users to find certain kinds of content including PPC and we have broad evidence on the extent of PPC-related harm. There are also differences in the framing of the duties in the Act which suggest that greater

[1004] Under section 29 of the Act. Please see Section 17, Search moderation for more detail on the relevant duties.

protections are expected to address PPC as they present a risk of harm to all children, irrespective of age.

22.20 For these reasons, our proposed service design measures primarily address the risks associated with children encountering PPC via search services. We have also included PC within the scope of Measure SD1, as we are aware of evidence that predictive search functionalities can result in user access and exposure to both PPC and PC. For example, predictive search suggestions can create a risk of children encountering search content that is abusive and hateful by clicking through harmful predictive search suggestions.[21]

22.21 The Act requires that service providers minimise the risk of children of any age encountering PPC, but the equivalent duty for other content that is harmful to children (i.e. PC and NDC) requires that the risk be minimised only for children in age groups judged to be at risk of harm from that content. We have inferred from these duties that greater protections are expected to address PPC as they present a risk of harm to all children, irrespective of age.

22.22 Search services allow users to search for content without being logged-in, making it difficult for search services to determine whether a user is under-18. We provisionally recommend that our measures apply to all users whether they are logged-in or logged-out. In making this recommendation, we have carefully considered the potential impact on service providers and users (particularly adult users), both in terms of user experience and the right to freedom of expression given that our measures will impact the use of search functionalities which can facilitate access to content. We consider our proposals to be proportionate for the reasons we set out below.

22.23 We therefore propose the following measures:

a) Measure SD1: We recommend that large general search services with predictive search functionalities should offer users a means to easily report predictive search suggestions which they believe can increase the risk of users encountering harmful PPC and PC. Providers of services likely to be accessed by children should review reported predictive search suggestions to determine if suggestions present a clear and logical risk of users encountering PPC and/or PC. If a risk is identified, service providers should take appropriate steps to ensure the reported suggestion is no longer recommended to any users.
b) Measure SD2: We recommend that large general search services likely to be accessed by children should provide crisis prevention information in response to known PPC-related search requests regarding suicide, self-harm and eating disorders. Crisis prevention information should be prominently displayed so that it is the first information users encounter in search results. It should include links to freely available, supportive information and helplines, provided by reputable mental health, suicide or eating disorder charities that hold relevant and accessible materials that are comprehensible and suitable in tone to all users, including children, in the UK.

22.24 We note that in addition to our measures set out here, Section 18, User reporting, additionally proposes to recommend that search services that are multi-risk for content harmful to children provide age-appropriate user support materials, ensuring that children understand the user tools and reporting and complaints functions on the service.

Measure SD1: Reporting and removal of predictive search suggestions that present a risk of users encountering PPC and PC.

Explanation of the measure

22.25 We propose to recommend that large general search services likely to be accessed by children operating predictive search functionalities should offer users a means to easily report predictive search suggestions that can increase the risk of user exposure to PPC and/or PC. Service providers should take the appropriate action to ensure reported suggestions presenting a clear and logical risk of users encountering PPC or PC are no longer recommended.

22.26 To effectively implement this measure, we recommend service providers:

a) Offer users the means to report predictive search suggestions that are believed to increase the risk of user exposure to PPC and PC;
b) review reported predictive search suggestions in line with its publicly available statement;[1005]
c) determine if the reported suggestion presents a clear and logical risk of users encountering PPC or PC; and
d) if the risk is identified, take appropriate steps to ensure that a reported suggestion is no longer recommended to any user.

22.27 We expect this measure to ensure that predictive search suggestions that are identified as potentially leading to PPC or PC are no longer presented to users. We propose to provide flexibility to services to decide the technical means by which they achieve this outcome. This is because we believe that services are best placed to determine what technical steps are most appropriate to achieve the desired outcome.

22.28 We consider it necessary to make specific recommendations for a targeted reporting mechanism for predictive search suggestions. This is because predictive search suggestions are not otherwise covered by the reporting and complaints duties imposed on search services in sections 31 and 32 of the Act (See Section 18, User reporting and complaints for more information on service reporting duties and related measures). This is because those reporting and complaints duties apply to 'search content' only. We do not consider predictive search suggestions to be 'search content' because they are generated by a separate underlying ranking algorithm that does not operate by means of the search engine.

22.29 Our measure seeks to address the risk of children encountering content that is harmful to children in the search results presented when a potentially harmful search suggestion is selected, not from encountering that content within the search suggestion itself.

22.30 We provisionally consider that, if services take steps to remove reported predictive search suggestions that present a clear and logical risk of directing users to PPC and PC, it will reduce the risk of other users, including children, being presented with these suggestions,

[1005] Please see Section 17, Search moderation, for more information on the relationship between publicly available statements and moderating content harmful to children.

and potentially of encountering PPC and PC. This is particularly the case compared to a counterfactual where predictive search suggestions remain unmoderated.

22.31 Our proposed measure does not restrict user ability to enter search requests or access search results**.** As such, we recognise that our measure may be less effective where children are intentionally searching for PPC and PC. We believe this measure will be most beneficial for children not actively searching for PPC and PC, but who may be predisposed to engage with PPC and PC if prompted or if exposed to suggestions which could lead to encountering PPC or PC.

Effectiveness of predictive search reporting at addressing risks to children

22.32 Current practice indicates that our measure is a technically feasible way for large general search services to reduce the risk of children encountering PPC and PC via the predictive search functionality.

22.33 We understand that it is current industry practice across large general search services, including Google Search and Microsoft Bing, to take steps to reduce the risk of harm posed by predictive search functionalities by identifying and preventing search suggestions that are harmful or violate their policies.

22.34 Both Google Search and Bing enable user complaints or reports related to predictive search suggestions. These reporting mechanisms may, in some cases, be accompanied by automated systems designed to prevent harmful predictive search suggestions being recommended to users. As explained above, however, we do not have evidence that current practices extend to all relevant categories of PPC or PC.

22.35 Some smaller services like DuckDuckGo have predictive search functionalities and complaints systems in place that enable it to receive reports to improve search results.[1006] Mojeek has an "autocomplete" function that shows previous search requests a user has entered but does not have a predictive search functionality that autocompletes user search requests or which suggests other search requests (i.e., from trending searches).[1007] Yahoo enables search predictions and allows users to report inappropriate predictions which are analysed against Yahoo's autocomplete policies.[1008]

22.36 As per Section 18, User reporting and complaints, research suggests that making reporting tools more prominent can make it easier for users to access and therefore increase the number of reports users make.

22.37 Google Search gives users the option to "report inappropriate predictions" on the search bar. This option is written in grey italics, found at the bottom of the suggestion box and in a font size smaller than the text which it surrounds.

22.38 Microsoft Bing has a "feedback" link at the bottom of the search results page that allows users to provide feedback on search "suggestions." Users can access the feedback link by clicking on the "settings and quick links" option at the side of the homepage, and then scroll half-way down to click on a "feedback" option; users are presented with a free-text reporting box where they can share feedback about what they "like", "dislike" or "suggest."

[1006] Duckduckgo, Settings (option to switch of autocomplete suggestions). [accessed 1 March 2024].
[1007] Mojeek, Appearance Settings (option to switch off Autocomplete). [accessed 1 March 2024].
[1008] Yahoo, About Yahoo Search Predictions [accessed 28 March 2024].

There is an additional link underneath the feedback box for a user to "report a concern" which takes the users to a page where they can specify reports about Bing. [1009]

22.39 We do not think the current practices of Google Search and Bing to report predictive search predictions are easy to find or access. To effectively reduce reporting barriers, we propose that reporting tools should be easy to find and easily accessible in relation to the predictive search suggestions themselves.

22.40 We considered whether to reframe this measure so that the provision of a reporting mechanism and the subsequent action applies only to predictive search results shown to users believed to be children by the service provider (see 'Other options considered' below). Research on user reporting behaviour, however, found that adults are more likely to report harmful content online than 13-17 year olds.[1010] Though related to experiences across different types of online service, this evidence highlights the advantage of our measure applying to all search service users, irrespective of age. We believe that ensuring all users can report predictive search suggestions will increase our measure's effectiveness as it will potentially increase the number of reports and draw the service provider's attention to potentially harmful predictive search suggestions. This would be the most effective and proportionate means to ensure that children are not recommended predictive search suggestions that present a clear and logical risk of encountering harmful content.

Effectiveness at minimising children's exposure to harmful content

22.41 In addition to reporting, Google Search relies on automated systems and enforcement teams to identify and remove problematic predictive search suggestions and closely related variations that violate the service's general and specific autocomplete policies.[1011] This includes predictions that contain dangerous (including self-harm), sexually explicit, harassing, hateful or terrorist content.[1012] Google Search also gives users the ability to turn off "trending" and "related" search recommendations.[1013]

22.42 Bing takes steps to ensure that users are not inadvertently led to "potentially harmful, offensive, or misleading content" via search suggestions. It states that it implements guardrails to prevent users from unexpectedly being exposed to potentially harmful or offensive content by using a combination of proactive and reactive algorithmic interventions.[1014]

22.43 Research indicates that Google Search's removal of specific antisemitic predictive search suggestions resulted in fewer search requests relating to this antisemitic proposition.[1015] This suggests that if services remove harmful predictive search suggestions, fewer users, including children, would encounter PPC and/or PC, because they would be less likely to be prompted to, and subsequently, search for it.

[1009] Ofcom desk research, conducted 19 March 2024, [web] Microsoft Bing, Report a Concern To Bing. [accessed 8 April 2024]
[1010] Ofcom, 2023. Online Nation.
[1011] [Web] Google, How Google autocomplete predictions work [accessed 12 December 2023]
[1012] Google, no date. Search Help, Content policies for Google Search [accessed 5 December 2023].
[1013] Google, no date. Google Search Help: Manage Google autocomplete predictions [accessed 5 December 2023].
[1014] Microsoft Bing, How Bing delivers search results [accessed 12 December 2023].
[1015] Antisemitism Policy Trust, 2019. Hidden Hate: What Google searches tell us about antisemitism today p.19.

22.44 As noted in the introduction, children may actively search for content which encourages, promotes or provides instructions for suicide, self-harm and eating disorders on social media platforms and be aware of codewords for PPC and PC-related content. Therefore, if search services take steps to remove predictive search suggestions which might lead to PPC and PC, alternative search terms could become less prominent. This effort would also reduce prompts to access content which may contain PPC and PC, thereby protecting children from those harms.

22.45 Cumulative exposure to harmful content is known to contribute to harm in children.[1016] We assess that children could be at risk of cumulative exposure if they are prompted or exposed to predictive search suggestions which could lead to encountering harmful content. Overall, we consider that this measure would effectively reduce the likelihood of children encountering PPC and PC by minimising prompts to PPC and PC via predictive search functionalities, and thereby reducing the quantity of harmful content accessed through search terms a child had not intended to search for. Our measure will expand on the harms covered in current practice to ensure that all relevant content categories of PPC and PC harms are in scope and users have the means to easily report predictive search suggestions that can lead to PPC and PC.

22.46 We acknowledge that most of our evidence relates to children's experiences searching for, and exposure to, PPC. However, we assess that where PC exists on the web, there is a risk that users can encounter it via general search services and that predictive search functions may direct users to this content. We logically assume that extending our measure to apply to PC will ensure the greatest impact on minimising children's exposure to PC and any resulting risk of harm.

22.47 We recognise that predictive search functionalities may also direct users to helpful resources and content; predictive search functions may, for instance, suggest search terms and direct users to content related to suicide prevention hotlines or useful resources and communities. Our measure does not propose that services turn off predictive search suggestions for search requests for specific terms or make any recommendations to impact the ranking of predictive search suggestions and search results. We have structured our measure in such a way to give flexibility to search services in how they identify risk and the steps they take once risk is identified to ensure that predictive search functions are not prevented from providing access to, or awareness of, support services.

Rights assessment

Freedom of expression

22.48 As explained in Section 2, Article 10 of the ECHR upholds the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. It is a qualified right, and Ofcom must exercise its duties under the Act in a way that does not restrict this right unless satisfied that it is necessary and proportionate to do so.

22.49 We do not consider that our proposed measure related to predictive search would have any material impact on users' (children's or adults') freedom of expression rights under Article 10. We acknowledge that the measure will have an impact on the availability of some search suggestions which are reported, and which services determine to present a clear and logical

[1016] Ofcom, 2022. Research into risk factors that may lead children to harm online.

risk of directing users towards search content that contains PPC and PC. The removal of the suggestion, however, would not prevent users, including an adult user, from inputting search requests or accessing search results through the service. Therefore, to the extent that our measure could amount to an interference with the rights of users (including adult users who are not specifically targeted by this measure), we consider it would be minimal. We note that while not the specific objective of our recommendation, it may have an ancillary positive impact on vulnerable adults at risk of suicide, self-harm and eating disorders, by reducing their risk of exposure to related suggestions and content.

22.50 For the reasons outlined above, we do not consider that actions taken by service providers in line with our proposed measure would have any material impact the freedom of expression rights of interested persons (i.e., website operators). This is because the website would remain discoverable via the search engine even where a predictive search suggestion that surfaces a URL is removed or otherwise obscured.

22.51 We also consider there to be limited impact on the freedom of expression rights of search services, whose right to impart information to users in the form of predictive search suggestions would be restricted, in that it would not restrict them from otherwise ensuring that users could still search for this content (to the extent not otherwise restricted by our Search moderation measures, see Section 17.

22.52 Overall, any impacts would be proportionate to the aims of the measure to minimise the risk of children encountering PPC and PC. The removal of search suggestions will make it less likely for users, particularly children, to be prompted to enter search terms and access content that present a risk of them going on to encounter PPC and PC by means of the service. It, therefore, seeks to address the potentially very severe harm that might arise to children online, in line with the legitimate aims of the Act.

22.53 Taking these reflections and the benefits to children into consideration, we consider that the impact of the proposed measure on the right to freedom of expression, above and beyond the requirements of the Act, is limited and proportionate.

Privacy

22.54 We believe the impact of this measure on the right to privacy is negligible. We acknowledge that user reports related to predictive search suggestions might generate new personal data or involve processing existing data for new purposes, if the service considered it appropriate to retain information about user reports of predictive search suggestions (for example, for prioritisation purposes). However, our measure does not suggest or require that service providers retain users' personal data. Where the reporting mechanisms put in place to implement our proposed measure involve personal data processing, services must comply with relevant data protection legislation, including applying appropriate safeguards to protect the rights of both children (who may require special consideration) and adults who may submit reports regarding predictive search suggestions.

22.55 Overall, we consider that the impact of our proposed measure on the privacy rights of users is minimal where services comply with relevant laws, and any interference is proportionate to the benefits to children, as compliance with this measure would aid in satisfying the provider's duties under the Act.

Impacts on services

22.56 There is a direct requirement in the Act that search services implement reporting and complaints systems to cover a wide range of topics; this requirement does not extend to complaints about predictive search suggestions. The costs of this proposed measure would relate to adapting reporting and complaints systems required by the Act to ensure the predictive search suggestions can be easily reported, and appropriate action taken in response. We would expect it to be relatively straightforward to do so. We anticipate this may require 20-40 days of software engineering time, along with an equal amount of non-software engineering time,[1017] which entails estimated one-off direct costs between £10,000 to £40,000. See Annex 12, for more information on our assessed labour costs.

22.57 We would also expect a service to incur ongoing maintenance costs to run the extended reporting and complaints system. If the annual maintenance costs were 25% of the implementation cost, we would expect these costs to be between £2,500-£10,000 per annum or higher if additional user interfaces need to be amended (i.e., Web browser, iOS, Android).

22.58 There will also be additional costs for the review of reported predictive search suggestions, which are likely to vary with the size and risk of the service. Larger services are likely to require a greater number of reviewers as we would expect them to receive a larger number of user complaints. Services will need to ensure that their system is set up effectively to handle complaints around harmful predictive search suggestions, and that teams responsible for reviewing predictive search suggestions can appropriately action complaints in line with services' publicly available statements and policies.

22.59 We believe the additional costs of this measure will be lower for search services who have implemented the predictive search measure as proposed in our Illegal Harms consultation. These would consist of a one-off cost to adapt the predictive search reporting system to allow for complaints of PPC and PC-related suggestions. For example, services that allow users to categorise their complaints will need to expand their current categorisation to include categories for PPC and PC harms. All services will also incur costs to ensure that once complaints are categorised, they are routed to the correct team responsible for reviewing reported predictive search suggestions. If costs were 50% of what they would otherwise be, we estimate the direct one-off costs would be £5,000 to £20,000 and ongoing costs would be £1,250 to £5,000 per year. Services that allow users to categorise complaints are likely to incur costs at the higher end of this range. In contrast, we do not expect that the costs to review reported predictive search suggestions will be materially lower for services that are also within scope of the proposed Illegal Harms measure; we would expect a greater number of reviewers to be required to deal with a larger number of complaints associated with the additional categories of harm.

22.60 We note that the costs for some service providers may be lower than our estimates where they already have part, or all, of the proposed measure in place to protect children. For example, Google Search and Microsoft Bing already have mechanisms for users to report predictive search suggestions. To the extent that existing policies do not adequately cover PPC and PC harms, there will still be an increase in costs to review user reports of predictive search suggestions.

[1017] This is consistent with our assumptions for illegal harms, from paragraph 22.23 of Ofcom 2023, Volume 4 Illegal harms consultation [accessed 5 December 2023].

Which providers we propose should implement this measure

22.61 The proposed measure mitigates the risk of children encountering content that is harmful to them in the search results via predictive search. This is achieved by user reporting mechanisms and, accordingly, service review of reported predictive search suggestions.

22.62 Our evidence highlighted in Section 7.10, Risk of harm to children on search services and Section 12, Service risk assessment guidance and risk profiles suggests that general search services pose a greater risk of harm as opposed to vertical search services. We consider that the benefits to children's safety of applying this measure to large general search services, including Google Search and Microsoft Bing, can be very material, as these services have large user bases (including many children), are more frequently used for searches, and use predictive search. We believe that the measure is proportionate for such services, given the expected benefits and the capacity for large search services to implement the measure.

22.63 We have minimal evidence about the current practices of smaller general search services, and the extent to which children use smaller general search services. We assess it would not be proportionate to extend this measure to smaller general services at this time as the costs are likely to be material for such services and the benefits of applying this measure to a service with limited reach are likely to be relatively small due to the smaller user base.

22.64 Therefore, we propose that this measure should apply to all large general search services likely to be accessed by children (regardless of risk level) with predictive search functionalities.

Other options considered

22.65 We considered additional options to address the risks of predictive search functionalities, including:

a) Services should use highly effective age assurance to target the predictive search protections outlined above at identified child users (thereby giving adult users a less restrictive experience),
b) Services should create a toggle to give users the option to easily turn on/off predictive search functionalities, predictive search switched off by default for users believed to be a child;[1018] and,
c) Services should proactively identify and prevent predictive search suggestions for PPC and PC-related terms.

22.66 We would need further evidence on the effectiveness and proportionality of these options and different associated aspects before recommending these measures.

22.67 We considered proposing service providers use age assurance to identify child from adult users – and where a user is identified as an adult, allow predictive search to include suggested terms that are identified as harmful to children. We do not, however, think it would be proportionate to recommend highly effective age assurance for search services, which is not required for search services under the children's safety duties in the Act (unlike some U2U services). We also took into account the nature of search services and how they operate.[1019] Search services allow users to search for content while logged-out or logged-in.

[1018] See Section 17, Search moderation, SM1, 'Who our measures apply to' for a definition of 'users believed to be a child'.
[1019] See Section 17, Search moderation, for further information on how search services operate.

If every user had to create an account and carryout age assurance, it would impose significant additional frictions on users before they could search for information. This could dissuade some users from using the service to access information and would have privacy impacts and cost implications. We acknowledge that this would also have commercial implications for services as it risks altering existing business models. We do not think the potential benefits to adult users of having access to a wider number of predictive search prompts would outweigh these impacts.

22.68 If a default 'off' setting for predictive search only applied to users believed to be a child, the efficacy of this option for children accessing search without the service treating them as a child (i.e., from an adult's account), would depend on children opting to switch-off predictive search. Evidence indicates that default settings are effective as users often do not change or move away from the default setting.[1020] As such, we provisionally consider that children that are not identified or logged-in are unlikely to switch off predictive search. We do not consider it proportionate to propose that predictive search is turned 'off' by default for all users including adults, as this would have significant impacts to rights to access information.

22.69 Some large general search service providers may already take proactive steps to minimise predictive search risks, rather than solely relying on user reporting. These might include use of automated systems to prevent the suggestion of harmful results and user controls to switch-off the functionality.[1021] However, we would need more evidence on the technical operations and underlying policies governing these efforts to make related recommendations in our Codes that are effective and do not disproportionally interfere with user rights or place disproportionate burdens on businesses. This is an area of development and we encourage service providers to share information in their consultation response about their automated technical approaches and underlying policies to moderate potentially harmful predictive search suggestions. This will help us understand the extent to which existing and developing methods could begin to address the risks associated with predictive search as a potential pathway to content that is harmful to children or illegal content.

22.70 As we learn more about children's search experience, children's experiences using predictive search functionalities, and the risks associated with predictive search functionalities, we may refine our approach.

Provisional conclusion

22.71 While there may be different approaches to address the risks associated with predictive search, given the harms this measure seeks to mitigate in respect of PPC and PC, as well as the risks of cumulative harm search services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCS E2 in Annex A8.

[1020] Competition and Markets Authority, 2022. Online Choice Architecture, how digital design can harm competition and consumers.
[1021] Google, no date. Manage Google autocomplete predictions. [accessed 20 February 2024]; Microsoft Support, no date. How Bing delivers search results. [accessed 20 February 2024].

Measure SD2: Provision of crisis prevention information in response to search requests related to suicide, self-harm and eating disorders.

Explanation of the measure

22.72 We propose to recommend that large general search services likely to be access by children employ means to provide crisis prevention resources in response to search requests related to suicide, self-harm and eating disorders.

22.73 Crisis prevention information should:

a) Be prominently displayed to users in search results,

b) Comprehensible and suitable in tone and content for as many users as possible, including children;[1022] and,

c) Include reliable and trustworthy professional support information and a helpline operated by reputable mental health, suicide prevention or eating disorder charities (as relevant) that are appropriate for children to use and accessible by children in the UK.[1023]

22.74 To implement this measure, services would need to work to understand the relevant terms and intent behind search requests related to suicide, self-harm and eating disorders, and deploy crisis prevention information in response to those requests. Though we believe that services are best positioned to determine which search terms or requests should be captured to generate crisis prevention information, and how to identify related search requests, we consider that, at a minimum, it would be appropriate for services to provide crisis prevention information in response to:

a) **General search requests for suicide, self-harm and eating disorders.** While we recognise this category is broad and could capture help or pop culture references, we consider that providing crisis prevention information in response to general search terms could provide timely assistance to search users, particularly users at earlier, speculative stages of searching for suicide, self-harm and eating disorder content.

b) **Search requests seeking specific, practical, or instructive information on suicide, self-harm and eating disorder methods.** Research sampling the search history of individuals hospitalised for suicidal thoughts and behaviours, identified that in 21% of cases, users had searched for information that matched their chosen suicide attempt method.[1024] **We therefore consider that crisis prevention efforts could have a strong impact on those users in a vulnerable state conducting a more specific category of search requests, and be effective in minimising their risk of encountering specific, instructive or practical information on suicide methods.**

[1022] See Section 19, Terms of service and publicly available statements for more information on which characteristics are important in determining whether provisions are clear and accessible to children.
[1023] See Measure US6, Section 21, User Support, for more information on 'appropriate support' for children.
[1024] Moon KC, Van Meter AR, Kirschenbaum MA, Ali A, Kane JM, Birnbaum ML., 2021, Internet Search Activity of Young People With Mood Disorders Who Are Hospitalized for Suicidal Thoughts and Behaviors: Qualitative Study of Google Search Activity. JMIR Ment Health. 8(10) [accessed 20 February 2024].

While this research relates to suicide, we assess that the benefits would extend to those searching for similar self-harm and eating disorder content.

22.75 To ensure effectiveness of our measure, we recommend that it apply to all logged-in and logged-out search users. Crisis prevention information should be comprehensible and suitable in tone for a wide range of users, including children, but services can provide this information in the format they believe to be most appropriate as long as it is prominently displayed so that it is the first information users encounter in search results.

Effectiveness at addressing risks to children

22.76 Section 18, User reporting and complaints cites studies that suggest that online self-help tools and support resources may be helpful for young people who have experienced suicidal feelings and other mental health concerns. We have explained the value of U2U crisis prevention efforts to mitigate the risk of harm to children posed by suicide, self-harm and eating disorder-related content. We believe this benefit applies to search services as well.

22.77 It is current industry practice to provide users with crisis prevention information (i.e., hotlines and other support resources) in response to search requests for harmful content, including searches that indicate a high intent of self-harm or suicide.[1025] Research from the Journal of Online Trust and Safety, for example, indicates that search requests in English for "suicide" and "kill yourself" on three different search engines were the most likely requests to surface direct support information.[1026]

22.78 Google Search aims to provide and improve the visibility of "authoritative information," such as hotlines or text support services, in search results in response search requests that indicate a high-risk of self-harm or suicide. It, for example, prominently displays the number for the National Suicide Prevention Lifeline in response to suicide-related requests; this takes precedence over other search results and is more prominent than an advertisement.[1027] Our research also suggests that Google Search provides the Samaritans' helpline number in response to "SOS situations," and the BEAT helpline number in response to requests related to eating disorders.[1028]

22.79 Though Bing displays crisis prevention information in response to suicide-related requests, we are not aware to what extent Bing provides this information in response to self-harm and eating disorder search requests. Bing has publicly stated that it may provide supplemental information, such as warnings and public service announcements, where it identifies search results may include harmful or misleading information, but it does not detail the type of search requests that surface this information, or the nature of the warnings surfaced.[1029]

[1025] Ofcom, 2023. Google Call for Evidence Response: Second Phase of Online Safety Regulation. [accessed 14 February 2024].

[1026] This research evaluates the results returned from both general suicide terms and terms related to specific suicide means across Google, Bing and DuckDuckGo. 2021, Journal of Online Trust and Safety. How Search Engines Handle Suicide Search Queries. [accessed 14 February 2024].

[1027] NY Times, 2010, 'Suicide' query prompts Google to offer hotline [accessed 22nd November 2023]; Google, no date. Find personal crisis information with Google Search. [accessed 14 February 2024].

[1028] As of 8th January 2024, qualitative desk research has demonstrated that Google search returns helpline numbers for Samaritans and BEAT in response to search queries containing terms associated with suicide or eating disorders. Samaritans also report that Google provides information to charities and crisis support lines in response to searches for content relating to suicide. Samaritans, 2022. Towards a Suicide-Safer Internet.

[1029] Bing, 2024. How Bing delivers search results. [accessed 16 February 2024]; Journal of Online Trust and Safety, 2021, How Search Engines Handle Suicide Queries, [accessed 28 November 2023].

22.80 DuckDuckGo, Ecosia, AOL and Yahoo also present crisis support information in response to user requests for suicide-related terms.[1030]

22.81 Search services' existing crisis prevention efforts are broadly welcomed by mental health and suicide prevention charities. In response to Google Search's launch of its crisis prevention efforts, Samaritans stated the importance of ensuring that "vulnerable and distressed people are steered towards safe spaces"[1031] given the volume of information that people can access online. Mental Health Innovations also indicated that 2% (30-40 people) of its daily conversations on the SHOUT support service were referred via signposts on Google Search and suggested that this demonstrates that "interventions such as this work to divert internet users" from potentially harmful searches.[1032]

22.82 We are conscious that the services in scope of this measure have a large user base, and that our measure could result in an unmanageable number of users being directed to crisis prevention resources and services, in particular, hotlines. While there is evidence that some viral user-generated content shared on a U2U service resulted in a spike of demand for Mental Health Innovations' SHOUT helpline Section 18, User reporting and complaints, we are not aware of concerns among third-party organisations that crisis prevention efforts by search services could lead to their support services becoming overwhelmed by additional demand generated by our measure.

22.83 Despite the different ways that services approach crisis prevention efforts, current industry practice suggests that this measure is a technically feasible way for services to minimise the risk of children encountering specific PPC related to suicide, self-harm and eating disorders.

22.84 Research confirms that the ranking of search results can have important implications for users. A discussion paper by the Competition and Markets Authority concludes that higher ranked items are more likely to be clicked on and chosen, and that users may perceive higher ranked results to be of better quality and relevance and reduce user effort to search through alternative options.[1033] Evidence suggests this is true for users searching, and accessing search results, for PPC harms, including suicide. A 2019 National Library of Medicine report, for example, found that higher ranked search results for suicide-related search requests, which were neutral and shown among anti-suicide pages, were more likely to be clicked on, concluding that efforts should be made to improve the visibility and ranking of suicide prevention webpages.[1034] This research suggests that users are more likely to access crisis prevention information if it is highly ranked, and that prominently displayed support information and helplines can help to minimise the risk of harm to users.

22.85 We suggest that crisis prevention information include links to support information and helplines operated by reliable and trustworthy professional support that is relevant and accessible to UK users. Services can provide this information in the format they believe to be most appropriate as long as it is prominently displayed so that it is the first information users encounter in search results, and comprehensible and suitable in tone and content for as

---

[1030] Ofcom 2023, Volume 4 Illegal harms consultation. [accessed 5 December 2023].
[1031] Samaritans, 2010, Google and Samaritans: new search feature to help people looking online for information about suicide [accessed 12 July 2023].
[1032] MHIUK response to 2023 Protection of Children Call for Evidence.
[1033] Competition and Markets Authority, 2022. Online Choice Architecture. How digital design can harm competition and consumers. [accessed 14 February 2024].
[1034] National Library of Medicine, 2019.Do Search Engine Helpline Notices Aid in Preventing Suicide? Analysis of Archival Data [accessed 28 November 2023].

many users as possible, including children. See Section 19, Terms of service and publicly available statements for details on characteristics important in determining whether provisions are clear and accessible to children.

22.86 Based on our understanding of current practice, we asses that to implement this option, services would need to generate keyword lists composed of suicide, self-harm, and eating disorder-related terms to detect and provide crisis prevention information in response to related search requests. Google Search, for example, takes keywords into consideration when reviewing content to determine whether it is policy-violating.[1035] Violative content includes PPC such as 'dangerous' content, including self-harm, and 'sexually explicit' content, including graphic sex acts.[1036] Google Search is working with machine learning and improving its AI models to automatically and more accurately detect a wider range of personal crisis searches, including topics such as suicide and abuse.[1037] We believe services are best positioned to determine what combination of terms should generate crisis prevention information.

Rights assessment

Freedom of expression

22.87 As explained in Volume 1, Section 2, Article 10 of the ECHR upholds the right to freedom of expression and encompasses the rights as set out in the 'Rights assessment' for Predictive Search (SD1) above.

22.88 By providing supportive information at a critical point in the user search journey, this measure adds friction to pathways to content that encourages, promotes, or provides instructions for suicide, self-harm and eating disorders, and seeks to address the potentially very severe harm that might arise to children who search for this content or might encounter it inadvertently. This is in line with the legitimate aims of the Act and the children's safety duties it imposes on search services. The measure will not impact the search results presented to users following the presentation of crisis prevention information. To the extent that this measure helps to prevent children from accessing suicide, self-harm and eating disorder content in search results, we consider that this will secure the objectives of the Act and is one of the least restrictive ways to secure them. Therefore, we consider any such impact on the rights of users, interested persons or search services to freedom of expression to be proportionate and justified to achieve the legitimate objectives of the Act.

22.89 We consider that there may be a limited impact on the freedom of expression rights of users of search services (including both children and adults), and those who impart beneficial and non-harmful content related to suicide, self-harm and eating disorders online (be this website operators or individual publishers), to the extent that they are also signposted to support if they search for this content. While the presentation of crisis prevention information may serve as friction in user journeys, users are not prevented from scrolling beyond this information and engaging with search results should they wish to do so.

[1035] Ofcom, 2024. Google response Qs 16-19, Call for evidence: Second phase of online safety regulation.
[1036] Google, no date. Content policies for Google Search. [accessed 1 March 2024].
[1037] Tech Crunch, no date. Google rolls out AI improvements to aid with Search Safety and 'personal crisis' queries. [accessed 1 March 2024]

22.90 Taking into consideration these assessments and the benefits to children, we consider that the impact of the proposed measure on the rights to freedom of expression, above and beyond the requirements of the Act, to be limited and proportionate.

Privacy

22.91 We recognise that depending on how service providers decide to implement the proposed measure, it could result in a greater or lesser impact on users' privacy rights under Article 8 of the ECHR as set out in Section 2.

22.92 The proposed measure does not specify that service providers should obtain or retain any specific types of personal data about individual users as part of their implementation of this measure. However, we recognise that, in particular, the analysis of search requests may involve processing personal data of the user conducting the search (although this may be no more than services ordinarily would process in delivering search results). Services which choose to process additional personal data in their analysis of search requests, or in any other process involved in implementing this measure, would need to comply with relevant data protection legislation. This would include applying appropriate safeguards to protect the rights of children (who may require special consideration) and adults who will be affected by this measure.

22.93 We therefore consider that the impact of the proposed measure to be very limited where services comply with relevant laws, and any interference with users' rights to privacy is necessary to secure that providers fulfil their children's safety duties under the Act, and proportionate to the benefits to children.

Impacts on services

22.94 We expect there will be costs to services associated with implementing and maintaining this measure.

22.95 While we are not prescribing how services identify search requests that should prompt crisis prevention resources, current industry practice suggests it likely involves some form of keyword detection. It is likely that services may need to adjust or implement systems that make use of a combination of technologies and inputs. For example, keyword lists supplied by specialist experts supplemented with machine learning and adjustments to existing search ranking algorithms to effectively identify and keep up to date on suicide, self-harm and eating disorder related requests, including specific language that individuals may use to avoid detection. Services will need to build or adapt existing systems that display crisis prevention information in response to identified requests/terms and run quality assurance to ensure the information surfaces correctly. This may involve services adapting their ranking system to prioritise resources.

22.96 We estimate that for services that do not currently have crisis prevention efforts in place, it could take an estimated 150-310 days of software engineering time and an equal amount of non-software engineering time to build a system that surfaces crisis prevention information in response to suicide, self-harm and eating disorder-related search requests. We estimate this could entail a one-off cost of around £74,000 to £308,000. See Annex 12, for more information on our assessed labour costs.

22.97 We also expect a service to incur ongoing maintenance costs to run and update the crisis prevention system. If the annual maintenance costs were 25% of the implementation cost, we expect these costs could be around £19,000 to £77,000. These costs could include

services ensuring keyword lists are updated to reflect changes in language associated with suicide, self-harm and eating disorders, and trends in new user methods to avoid detection. Services would also need run quality assurance for any new terms to ensure that the crisis prevention information continues to surface correctly.

22.98 Service providers that already implement our related Illegal Harms measure will have to expand their crisis prevention to cover requests relating to self-harm and eating disorders. If providers use keyword detection to identify harmful requests, this will involve expanding their keywords lists to cover self-harm and eating disorder-related terms and ensuring that the crisis prevention system responds correctly to these new terms. If these costs represented 50% of the implementation costs estimated above, we expect the incremental cost (over and above the costs estimated for the IH measure) to expand the system to self-harm and eating disorder related requests could be between £37,000 to £154,000. We assume this could result in annual maintenance costs of around £9,500 to £38,500. As above, this will involve quality assurance and keeping keyword lists related to self-harm and eating disorders up to date.

22.99 To implement this measure, service providers will also need to identify reputable charities with hotlines and resources related to suicide, self-harm and eating disorders. This may entail some research, which we envisage would result in a small cost as the number of possible organisations which are relevant and accessible to UK users is relatively small. We also expect there to be ongoing costs to ensure hotline and resource information is up to date.

22.100 We recognise that the charities selected to provide crisis prevention information may incur additional costs from their helplines and websites receiving additional traffic. We believe this risk may be limited for the reasons outlined in paragraphs 1.78 – 1.80, including because the services currently in scope of this measure already have forms of crisis prevention in place that link to the websites and hotlines of reputable UK-based charities working in the suicide and mental health (including eating disorder) space. Any additional traffic that comes from this measure will therefore be limited to where services' current crisis prevention resources don't sufficiently cover self-harm, suicide and eating disorders.

22.101 We note that the costs for some service providers may be lower than our estimates where they already have part, or all, of the proposed measure in place to protect children. For example, we are aware that many general search services already have crisis prevention mechanisms in place for some harmful search requests. If its policies already sufficiently cover suicide, self-harm and eating disorder requests, there may be no incremental costs provided it does not want to withdraw these measures. To the extent existing policies do not adequately cover these requests, there will be an increased cost to ensure crisis prevention mechanisms are extended to cover all suicide, self-harm and eating disorder-based requests.

Other options considered

22.102 We considered framing the measure to propose that service providers do not need to direct crisis prevention information to users believed to be adults. We believe that this framing would reduce the measure's effectiveness to mitigate the risk of harm to children. As previously explained, search services allow users to access search while logged-out or logged-in. By proposing that crisis prevention information is not presented to users believed to be adults, there is a risk that the same information would not be provided to logged-out

children, or those children who have not provided an accurate data of birth at sign-up or are accessing search from an adult's account.

22.103 We also considered whether to recommend highly effective age assurance to services to determine the age of a user. However, unlike for U2U services, the Act does not suggest that use of access controls is a type of measure that search services should consider using to meet their children's safety duties to minimise the risk of children encountering harmful content.[41] To help search services meet this duty, we do not believe it is currently proportionate to structure the measure to rely on highly effective age assurance to target the crisis prevention measure at children. As such, we believe that our proposed recommendation that the measure apply to all users effectively helps search services meet their children's safety duties.

22.104 We considered recommending that large general search services provide crisis prevention information in response to user search requests for other types of PPC harms (that is, pornography) and PC harms. We acknowledge there are organisations providing supportive information to users for pornography and PC harms, such as bullying or hateful content directed at protected characteristics. We are currently unaware, however, of any dedicated charities that serve to support children that may be undergoing crises linked with viewing pornography, in particular. Further to this, we may consider the role of supportive resources for harms such as intimate image abuse and controlling and coercive behaviour in our forthcoming guidance on protecting women and girls.

Which providers we propose should implement this measure

22.105 We believe this measure can materially reduce harm by intervening at a crucial point of the user search journey and providing relevant resources to users, including children, who may be at risk of severe harm.

22.106 We consider this measure proportionate for large general search services likely to be accessed by children given the evidence to suggest crisis prevention can disrupt harmful user journeys, discourage future harmful searches, and discourage user engagement with otherwise harmful suicide, self-harm and eating disorder-related content. The large user bases of these services – including many children – suggests that the benefits of applying this measure to these services are likely to be material, and we believe that the costs are likely to be proportionate for such services.

22.107 We further believe that as proposed, our measure is a technically feasible way for services to meet their children's safety duties as we note that many large general search services already have in place at least some form of the crisis prevention measure we are recommending. We expect that operating a large general search service requires systems to effectively categorise search requests, which will decrease the cost of implementing crisis prevention if not already in place.

22.108 At this time, we do not propose to extend this measure to smaller search services. Smaller services have a lower reach, and we do not currently have sufficient evidence to suggest that the material cost of this measure would be proportionate for such services given that fewer search pathways of children towards PPC would be disrupted.

22.109 As noted, however, some smaller search services, including DuckDuckGo, Ecosia, AOL, and Yahoo, already voluntarily provide crisis prevention information for suicide-related search queries. Though we do not currently consider smaller services in scope, we encourage them

to continue to provide crisis prevention information where they currently do so given the highlighted benefits.

22.110 We therefore propose that this measure should apply to all large general search services likely to be accessed by children (regardless of risk level).

Provisional conclusion

22.111 Given the harms this measure seeks to mitigate in respect of suicide, self-harm and eating disorder content, as well as the cumulative harm search services pose to children, we consider this measure appropriate and proportionate to recommend for inclusion in the Children's Safety Codes. For the draft legal text for this measure, please see PCS E3 in Annex A8.

23. Combined Impact Assessment

In the preceding sections we have set out our assessment of the impacts of each proposed measure. In this section, we set out our assessment of the combined impact of our package of proposed measures. At the heart of our assessment is the extent to which our measures can reduce the risks that children face when using different kinds of services. We also consider potential adverse impacts on users and services, ultimately aiming to ensure that the measures – both individually and as a package – will protect children online without unduly affecting user rights or undermining innovation and investment in high-quality online services for UK citizens.

Overall, our codes place more demanding expectations on services that pose greater risk of harm to children, even if they are smaller services, because this is where measures have the greatest potential to support safer experiences for children online.

Our provisional conclusion is that the measures as a whole are proportionate. To reach this view, we have considered implications for different kinds of services:

- *Smaller, low-risk services*. These services are in scope of a limited set of core measures recommended for all services, largely based on specific duties in the Act. The proposed measures for these services represent our view of the baseline measures required for all services to comply.

- *Smaller services with medium or high risks*. In addition to the core measures, smaller risky services will also be in scope of various additional measures that target specific risks and functionalities – including User Support and Recommender System measures – strengthening the protection of children from specific harms. These targeted measures are complemented by more cross-cutting measures that help protect children from *all* harms. In particular, for services with a relatively complex risk environment (posing risks of multiple kinds of content harmful to children), we recommend more sophisticated governance and content moderation systems and processes. While we have sought to give services flexibility in how to implement measures according to their risk and context, the total cost of this package of measures may be high for services that have several risks. Some small or micro businesses could struggle to implement these measures, potentially leading to degradation of user experience or even withdrawal of services from the UK. However, on balance, we consider the proposed measures are proportionate given the risks posed by these services.

- *Large services*. In addition to the core measures, we recommend some measures – including certain governance and content moderation measures – for all large services, regardless of risk. Large services that meet relevant risk criteria need to apply the same targeted measures as those that apply to smaller risky services, and in a few specific cases we recommend further measures for large risky services only. In aggregate, this results in a potentially costly set of proposed measures for large services. We consider this proportionate given the significant scope to reduce the risk of harm that they pose for the many UK children that use them, and the ability of large services to absorb the costs of the package of measures we propose.

- *Services whose principal purpose is the hosting or dissemination of PPC or PC*. For these services, we recommend highly effective age assurance to prevent access to the service by children. While the cost of implementation can be substantial, we consider this

proportionate and crucial to reduce the severe and inherent risk to children from these services.

Consultation Questions

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?

Introduction

23.1 In the preceding sections we have set out our assessment of the impacts of each proposed measure. In this section, we set out our assessment of the combined impact of the package of recommended measures that we are proposing and explain why, seen in the round, we consider them to be proportionate.

23.2 In assessing combined impacts, we apply our impact assessment framework, as described in our Framework for Codes (Section 14). At the heart of our assessment is the extent to which our package of measures can reduce the risks that children face when using regulated services. This allows us to identify which measures are most effective at protecting children and to target those measures towards services where children face the greatest risks, also taking into account any potential adverse impacts of our recommendations.

23.3 Protecting children online in line with the Act will inevitably entail material costs to business and could even lead to some services exiting the UK. However, while striving to ensure that the package of measures delivers a higher standard of protection for children than for adults, we are realistic that it is not possible for these measures to eliminate fully the risk of harm to children. We are mindful of potential adverse effects, including where disproportionately high costs to businesses could reduce innovation, investment, competition and market entry, which would ultimately harm users who benefit from, or rely on, the many diverse services in scope of the Act. Therefore, our combined impact assessment aims to ensure that the package of proposed measures will be effective in protecting children online, without unduly affecting user rights or undermining innovation and investment in high-quality online services for UK users.

23.4 Considering impacts cumulatively builds on the assessments provided in previous sections, which focus on measures individually. It allows us to consider the combined benefits and costs of measures, which may far exceed the benefit and cost of any individual measure. For example, if certain measures complement one another, then the incremental effect of each measure in reducing risk of harm may be substantial. On the other hand, if several measures all target the same specific risk factor, the incremental benefit of each measure might be limited.

23.5 The Act requires that we have regard to specific principles, including the proportionality of measures for different services, which may depend on the size and capacity of the services in question.[1038] Therefore, in this section we consider how the package of measures may affect different kinds of services.

23.6 Our assessment reflects that the kinds of services in scope of each proposed measure varies. As explained in the Framework for Codes (Section 14), while some measures apply to all services, in other cases we use various criteria to define which services are in scope of each measure:

a) **Risk**. Some measures target **specific risks**, where services have one or more medium or high risk for specific kinds of content that are relevant to each measure.[1039] We also set

[1038] Schedule 4 of the Act.

[1039] For example, Recommender System Measure RS1 applies to services with medium or high risk for at least one kind of PPC (i.e., at least one of pornography, suicide, self-harm or eating disorder content). Other measures define alternative combinations of kinds of content, e.g., User Support Measure US2 applies to services with medium or high risk for at least one of bullying, abuse and hate, and violent content.

additional expectations on services that are **multi-risk**, meaning that they have medium or high risk for at least two kinds of content harmful to children,[1040] whatever those kinds of content may be. These are more general measures that can contribute to harm reduction in respect of all kinds of content harmful to children, where a service operates in a more complex risk environment with multiple interdependent risks to manage.

b) **Size of the user base**. We define **large services** as those with an average user base greater than 7 million monthly UK users. We refer to services below this threshold as **smaller services**.

c) **Functionalities**. Some proposed measures are dependent on whether a service has certain functionalities (e.g. recommender systems) or other characteristics (e.g. uses volunteer moderators).

23.7 We are required to consider impacts on small and micro businesses in particular.[1041] Such businesses may have relatively limited capacity to implement measures. By assessing impacts on these businesses, we aim to ensure that costly measures are proportionate given their potential to improve child safety online, recognising where there may be some adverse impacts (for example, if small and micro businesses reduce investment or stop operating due to the cost of proposed measures).

23.8 As explained in the Framework for Codes (Section 14), we adopt commonly used definitions of **small and micro businesses**, based on having 10-49 and 1-9 full-time employees respectively. This is distinct to our definitions of services as noted above, though these will overlap. Small or micro businesses are likely to reach fewer than 7 million monthly UK users, and therefore qualify as smaller services. We recognise that the Act also applies to non-commercial entities, and where such entities are small, we consider that they may face similar impacts to those on small or micro businesses.

23.9 Where we propose measures for services regardless of their size, our assessment considers whether our proposed package of measures is proportionate for services provided by small and micro businesses. If they are, we believe that these measures will also be proportionate for larger businesses. We separate our assessment according to the following categories of smaller services for the purposes of this section:

a) A core set of measures recommended for all services, including low-risk services.

b) Targeted measures for any services that meet specific criteria, including specific risks, functionalities or other characteristics. These should be implemented in addition to the core measures recommended for all services.

c) Measures recommended for all multi-risk services, which should be implemented alongside the core measures and any applicable targeted measures.

23.10 We also consider impact where measures do differentiate based on size. We assess these separately, distinguishing between:

a) Measures recommended for large services regardless of risk. We recommend that all large services, even if low-risk, should implement additional measures – primarily related to governance and content moderation – as well as the core measures recommended for all services.

[1040] I.e., medium or high risk for at least two kinds of content among the four kinds of PPC, eight kinds of PC and any applicable kinds of NDC.

[1041] See Section 7 of the Communications Act, as amended by Section 93(4) of the Online Safety Act.

b) Measures recommended for large services if they meet additional criteria (including being multi-risk or posing certain specific risks). These should be implemented alongside the core measures recommended for all services, the measures recommended for all large services, and any other applicable measures that are recommended for certain services regardless of size, as described above.

23.11 It should be noted that the following sub-sections deliberately do not cover our age assurance Measures AA1 and AA2. These measures recommend that services whose principal purpose is to host or disseminate PPC or PC should implement highly effective age assurance to prevent children from accessing the entire service. Implementing this measure is intended to achieve the result that such services would no longer be likely to be accessed by children, meaning that they would be expected to become out of scope of the children's risk assessment and safety duties.[1042] Such a service would therefore not be expected to implement these measures in combination with other measures, so we do not consider any cumulative impacts. As set out in the age assurance section, we consider these measures proportionate because of the benefit from preventing children from accessing services that pose unacceptable risks to them, acknowledging that this could affect the commercial viability of some services, as well as people's ability to access legal content.

Impact of proposed measures recommended for services of all sizes

23.12 This section considers measures that apply to services regardless of their size, including smaller services (those with fewer than 7 million monthly UK users). In line with our duties, our analysis in this section includes an assessment of the impacts on small and micro businesses in particular. We expect that the vast majority of services provided by small and micro businesses will be smaller services, whereas the operation of large services is typically expected to require greater resources than those available to a small or micro business.

23.13 In this section we consider, in turn:

a) Measures recommended for all services. (Applies to U2U and Search services)
b) Measures recommended for services that meet specific additional criteria, including specific risks, functionalities or other characteristics. (Applies to U2U services only)
c) Measures recommended for all multi-risk services. (Applies to U2U and Search services)

Proposed core measures recommended for all services

23.14 The core measures below are recommended for all services, regardless of size, risk or any other criteria. This would therefore include low-risk[1043] services provided by small and micro businesses, among others.

[1042] Where a service implements highly effective age assurance, it can carry out a new children's access assessment to determine whether it is out of scope. Services not likely to be accessed by children must still comply with the duties about children's access assessments, which include a requirement to carry out a children's access assessment every year and sometimes more frequently. See Volume 2.
[1043] We refer to a service as 'low risk' if it does not have medium or high risk for any kind of content harmful to children.

Table 23.1: Summary of proposed measures recommended for all services

No.	Description of proposed measure	Services we propose this will apply to
GA2	Name a person accountable to most senior governance body for compliance with children's safety duties.	All user-to-user and search services
CM1	Content moderation systems and processes designed to swiftly take action against content harmful to children.	All user-to-user services
SM1	Have moderation systems and processes in place to take appropriate action on Primary Priority Content, Priority Content and Non-designated Content	All search services
UR1	Have complaints processes which enable users to make relevant complaints for services likely to be accessed by children.	All user-to-user and search services
UR2	Have easy to access and use, and transparent complaints systems.	All user-to-user and search services
UR3	Acknowledge receipt of complaints with indicative timeframe and information on resolution.	All user-to-user and search services
UR4	User-to-user services take appropriate action in response to each complaint.	All user-to-user services[1044]
UR5	Search services take appropriate action in response to each complaint.	All search services[1045]
TS1	Terms and statements regarding the protection of children should contain all information mandated by the Act.	All user-to-user and search services
TS2	Terms and statements regarding the protection of children should be clear and accessible.	All user-to-user and search services

User-to-user services

23.15 All U2U services in scope of the Act will need to take some measures to meet the important new duties that the Act places on them.

23.16 The measures we propose recommending for all services, even if low-risk and operated by small and micro businesses, can be divided into two groups. The first group directly reflect specific duties in the Act.[1046] This applies to our proposed measures for terms of service or publicly available statements and most of our proposed measures related to reporting and complaints. We have limited discretion over how this first group of measures should apply as the requirements in the Act are already very specific.

23.17 This first group of proposed measures may require material changes to systems, processes and service design. This would be the case, for example, for any kind of service that does not currently address risks to children in the terms of service or does not have a complaint handling function. Our impact assessment does not consider these impacts in detail, as Ofcom is not making decisions about those specific duties. We are concerned with how our measures meet those specific duties, and we consider our proposed measures set out a

[1044] Note that some additional steps are recommended for certain services based on size and risk criteria and whether they are in scope of age assurance measures. See Section 18, Volume 5 on User reporting and complaints for more information.
[1045] Note that some additional steps are recommended for certain services based on size and risk criteria. See Section 18, Volume 5 on User reporting and complaints for more information.
[1046] For example, Sections 20 (2) and 31 (2) of the Act place duties on providers of U2U and search services to operate systems and processes that allow people in the UK to easily report content harmful to children; Sections 12(9), 12(11)(a), 12(12), 21(3), 29(5), 29(7) and 32(3) set out requirements for providers to explain in their terms or statement the details of certain provisions taken to keep children safe on their service.

reasonable way of meeting those requirements, giving services considerable flexibility in how they chose to do that where appropriate.

23.18 Our impact assessment is focused more on measures related to duties in the Act which are less specific.[1047] We have more discretion over what these measures should cover and who they should apply to. For U2U services, we propose only three such measures for all services, even if they are small and low-risk:

a) A named person is accountable to the most senior governance forum for compliance with child safety duties, reporting and complaints duties. We consider the costs of this will be small or even negligible for smaller, low-risk services, but we believe the measure has potential benefits for all services. For example, naming an accountable person at an early stage could help to manage risk more effectively as a service evolves, including where new risks emerge.
b) Indicative timeframes for considering complaints should be sent to complainants. Although not explicitly required for all services by the Act, we consider that this has the potential to support user trust in the reporting process (which is a known issue among children and a barrier to reporting harmful content), at minimal cost even for small or micro businesses.
c) Content moderation systems and processes to swiftly take appropriate action in relation to harmful content. Although the Act does not explicitly require specific content moderation measures from all services – only requiring this 'if it is proportionate to do so' – in practice we consider that the proposed measure represents minimum steps that any service would need to take to comply with the Act. The measure is flexible and avoids undue costs for smaller, low-risk services that receive few or no user complaints.

23.19 In summary, small and micro businesses that operate a low-risk U2U services will only be in scope of a limited set of measures as outlined above. Most of these reflect specific Act requirements, and we consider that the measures have the potential to materially improve the safety of children while being manageable even for small and micro businesses to implement.

23.20 We also consider that most of these measures will overlap substantially with similar measures proposed in our Illegal Harms Consultation. The duties related to illegal content and content harmful to children are separate, and we expect significant incremental benefits from extending measures to cover content harmful to children as well as illegal content. These overlaps may also reduce the cost of implementing some measures proposed in the current consultation. For example, a service may choose to use the same complaints process for user complaints related to illegal content and content harmful to children.

Search services

23.21 All search services in scope of the Act will also need to take some measures to meet the significant and important new duties that the Act places on them.

23.22 For search services there are three measures recommended for all services regardless of size and risk, where we exercise a degree of discretion with a similar rationale to the U2U equivalents discussed above. We recommend that all search services should have:

[1047] For example, the duties in Sections 12(2), 12(3) of the Act which specify broader requirements relating to mitigating risk of harm to children and preventing or protecting children from encountering harmful content.

a) A named person accountable to the most senior governance forum for compliance with child safety duties, reporting and complaints duties;

b) Indicative timeframes for considering complaints should be sent to complainants; and

c) A moderation system and processes in place to take appropriate action on identified PPC and PC.

23.23 We believe the combined impact of these additional proposed measures on small and micro businesses that are low-risk for all harms would be limited. We expect that naming an accountable person would not substantially increase the person's workload in the case of a small, low-risk service, as that service is likely to have low volumes of harmful content and receives very few complaints. We consider that small and micro businesses would generally have the technical and financial capacity to undertake these measures.

23.24 Overall, we consider the measures that apply to all services to be proportionate and important in providing a baseline level of protection for children, upon which additional measures will build where services meet the relevant criteria, as explained in the following sub-sections.

Additional proposed measures recommended for services that meet specific risk criteria

23.25 In addition to the proposed core measures for all services, as summarised in the previous sub-section, the measures below are recommended for U2U services that meet additional risk criteria, regardless of size. The measures would therefore apply to any services provided by small and micro businesses that meet the relevant criteria. These vary for each measure and capture services with:

a) High or medium risk for one or more specific kinds of content; and/or
b) Relevant functionalities or other characteristics, such as using recommender systems or allowing some kinds of content harmful to children on the service.

Table 23.2: Summary of additional proposed measures recommended for services that meet specific criteria

No.	Description of proposed measure	Services we propose this will apply to
AA3	Use highly effective age assurance to ensure children are prevented from encountering Primary Priority Content identified on the service.	User-to-user services: • Whose principal purpose is not the hosting or the dissemination of one or more kinds of Primary Priority Content, and • Which do not prohibit one or more kinds of Primary Priority Content.
AA4	Use highly effective age assurance to ensure children are protected from encountering Priority Content identified on the service.	User-to-user services: • Whose principal purpose is not the hosting or the dissemination of one or more kinds of Priority Content • Which do not prohibit one or more kinds of Priority Content, and • Are medium or high risk for one or more kinds of Priority Content that they do not prohibit.

No.	Description of proposed measure	Services we propose this will apply to
AA5	Use highly effective age assurance to apply relevant recommender system measures in the Code to children.	User-to-user services that: • Are medium or high risk for one or more kinds of Primary Priority Content, and • Operate a content recommender system.
AA6	Use highly effective age assurance to apply relevant recommender system measures in the Code to children.	User-to-user services that: • Are medium or high risk for one or more kinds of relevant Priority Content (excluding bullying),[1048] and • Operate a content recommender system.
RS1	Ensure that content likely to be Primary Priority Content is not recommended to children.	User-to-user services that: • Operate a content recommender system, and • Are medium or high risk for at least one kind of Primary Priority Content.
RS2	Ensure that content likely to be Priority Content is reduced in prominence on children's recommender feeds.	User-to-user services that: • Operate a content recommender system, and • Are medium or high risk for at least one kind of Priority Content (excluding bullying).[1049]
US1	Provide children with an option to accept or decline an invite to a group chat.	User-to-user services that: • Have group chats, and • Are medium or high risk of one or more of: pornographic content, eating disorder content, bullying content, abuse and hate content[1050] and violent content.[1051]
US2	Provide children with the option to block and mute other users' accounts.	User-to-user services that: • Have user profiles and certain user communication functionalities,[1052] and • Are medium or high risk of one of more of: bullying content, abuse and hate content and violent content.
US3	Provide children with the option to disable comments on their own posts.	User-to-user services that: • Have comment functionalities, and • Are medium or high risk of one or more of: bullying content, abuse and hate content and violent content.

[1048] We are also minded to extend this measure for two potential kinds of Non-designated Content. See Section 15, Volume 5 on Age Assurance, for more information.
[1049] We are also minded to extend this measure for two potential kinds of Non-designated Content. See Section 20, Volume 5 on Recommender systems on U2U services for more information.
[1050] We use 'abuse and hate' content to refer to the two kinds of content defined in the Act in sections 62(2) and 62(3). A service is considered to have medium or high risk for abuse and hate content if it has medium or high risk for at least one of the two kinds of content defined in the Act in sections 62(2) and 62(3).
[1051] We use 'violent content' to refer to the three kinds of content defined in the Act in sections 62(4), 62(6) and 62(7). A service is considered to have medium or high risk for violent content if it has medium or high risk for at least one of the three kinds of content defined in the Act in sections 62(4), 62(6) and 62(7).
[1052] Please refer to Section 21, Volume 5 on User Support for more information on the functionalities that are applicable to Measure US4.

No.	Description of proposed measure	Services we propose this will apply to
US5	Signpost children to support at key points in the user journey.	**Intervention point 1 – when children report content** User-to-user services that are medium or high risk of one or more of: suicide content, self-harm content, eating disorder content, or bullying content. **Intervention point 3 – when children search for harmful content:** User-to-user services that: • Have user-generated content searching; • Are medium or high risk for one or more of: suicide content, self-harm content, eating disorder content, or bullying content; and • Have measures that enable them to become aware of when a user searches using suicide, self-harm or eating disorder related search terms.

23.26 As part of our combined assessment, we have considered that each of these measures only applies where services pose specific risks to children that are relevant to the measure, meaning services only incur costs to implement these measures where they can materially reduce the risk to children.

23.27 Our assessment of the benefits of these measures discussed in this sub-section also reflects that they are targeted at a range of different specific risk factors (including functionalities in many cases). Therefore, these targeted measures can each deliver benefits that are distinct from other measures, such as those related to governance and content moderation that address internal systems and processes rather than end-user functionalities. We also consider that these targeted measures will have significant incremental benefits over the measures discussed in previous sub-sections, which are more cross-cutting in nature.

23.28 With respect to specific measures, we expect our recommender system measures to deliver significant incremental benefits over other measures by directly addressing a functionality that has been shown to play a key role in amplifying exposure to content harmful to children. We recognise that the measures may involve substantial incremental costs for services that choose to operate a recommender system. This may include the costs of changes to these systems, as well as implementing highly effective age assurance under the age assurance measures, AA5 and AA6, which are linked to recommender systems. However, we consider that costs will typically depend on the complexity of recommender systems, with smaller services more likely to use simpler recommender systems, or not use them at all. We also expect that the costs of conducting age checks via a third-party provider will largely scale with the size of the user base and would therefore be lower for smaller services.

23.29 Our other age assurance measures[1053] are also expected to deliver important benefits by enabling children to be identified accurately and provided with safer experiences:

[1053] Note that services in scope of any of age assurance Measures AA3 to AA6 are also recommended to take certain steps in relation to handling complaints about incorrect age assessments, under the proposed reporting and complaints measure UR4(c). In our combined impact assessment, we have also considered the impacts of these additional steps, which we consider proportionate for the reasons set out in Section 18, Volume 5 on User reporting and complaints.

a) For services that do not prohibit all kinds of PPC (but do not have PPC as their principal purpose), we recommend the use of highly effective age assurance to prevent children from encountering identified PPC. This reflects a specific requirement in the Act itself over which we have limited discretion.
b) For services that do not prohibit all kinds of PC (but do not have PC as their principal purpose) and are medium or high risk for at least one kind of PC that they do not prohibit, we similarly recommend the use of highly effective age assurance to protect children from encountering identified PC. While we expect that the costs of this can be significant, we have also designed the measure flexibly – for instance, services may choose to apply highly effective age assurance only to users who specifically request access to PC. We consider that the costs of the measure are likely to scale with benefits to children.

23.30 Overall, we consider our measures related to group chat invites, disabling comments and blocking/muting users to provide important benefits by providing children with tools to manage their interactions with other users online. This gives them more control to avoid potential harm by declining undesired engagement with other users or ending engagement with users when it becomes harmful. Some of these harms are typically more challenging for providers to address solely through alternative means such as content moderation (e.g., bullying, abuse and hate content can be very context-dependent and can occur on private communication channels). Therefore, we consider these measures proportionate given their potential to strengthen children's safety, taking into account the material costs they may entail for services and potential added friction for users.

23.31 A further measure recommends signposting users to support resources when they search for or report harmful content. It complements other measures – that primarily aim to reduce the incidence of harmful content being encountered by children – by intervening at crucial parts of the user journey, such as immediately before or after a user may experience harm. This would help reduce the impact of harm in those cases where it does still occur. This measure is designed flexibly and expected to allow small and micro businesses to implement it at a low cost.

23.32 We acknowledge that these measures may add significantly to the costs that smaller risky businesses would face due to the recommended core measures for all services, discussed in the previous sub-section, as well as the relevant targeted measures according to a service's criteria as discussed in this section. We believe it is unlikely that services provided by small or micro businesses will meet all, or even most, of these criteria. For instance, a service that has a recommender system, volunteer moderators, a group messaging functionality and commenting functionality is inherently more likely to be a more complex service, and therefore more likely to be operated by a business with more resources.

23.33 Nonetheless, we cannot rule out that the targeted measures proposed in this section would add material costs for some small and micro-businesses. Where measures are linked to functionalities – which may broadly benefit users but may also lead to harm to children in certain cases – it is possible that some services with limited resources might be discouraged from offering those functionalities to users. In more extreme cases, the overall cost might even discourage some businesses from making their services available to UK users, or prevent some businesses from continuing to operate, which could impact users.

23.34 Equally, cumulative costs will be higher where services are in scope of many of these measures, but we consider such services will typically be relatively complex (e.g., with many

functionalities) and more likely to have sufficient capacity to implement the measures. Costs should also scale with benefits, as services with many relevant functionalities and risks for relevant kinds of content would generally pose greater risk of harm, absent appropriate protections.

23.35 Our provisional view, based on the factors outlined above, is that the combined impact of the measures in this sub-section, alongside the core measures for all services, remains proportionate. While we have considered the potentially adverse effects on users and small businesses, we believe there are large potential benefits in terms of reducing harm to children.

Additional proposed measures recommended for multi-risk services

23.36 The measures below are recommended for multi-risk services, regardless of size. This would therefore include multi-risk services provided by small and micro businesses, among others. Such services would also be in scope of core measures for all services (as discussed under 'Proposed measures recommended for all services'), as well as any of the targeted measures in the previous sub-section ('Additional proposed measures recommended for services that meet specific criteria') if they meet the relevant criteria.

Table 23.3: Summary of additional proposed measures recommended for multi-risk services

No.	Description of proposed measure	Services we propose this will apply to
GA3	Written statements of responsibility for senior members who make decisions relating to management of child safety risks.	Search and user-to-user services that are either: • multi-risk for content harmful to children; or • large user-to-user services; or • large general search services.
GA5	Track unusual increases or new kinds of Primary-Priority Content, Priority Content, and Non-designated Content on a service.	Search and user-to-user services that are either: • multi-risk for content harmful to children; or • large user-to-user services; or • large general search services.
GA6	Have a Code of Conduct that sets standards for employees around protecting children.	Search and user-to-user services that are either: • multi-risk for content harmful to children; or • large user-to-user services; or • large general search services.
GA7	Ensure staff involved in the design and operational management of service are sufficiently trained in approach to compliance with children's safety duties.	Search and user-to-user services that are either: • multi-risk for content harmful to children; or • large user-to-user services; or • large general search services.
CM2	Set internal content policies.	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.
CM3	Set performance targets for content moderation function.	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.
CM4	Have and apply policies on prioritisation of content for review.	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.

No.	Description of proposed measure	Services we propose this will apply to
CM5	Ensure content moderation functions are well-resourced.	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.
CM6	Ensure content moderation teams are appropriately trained.	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.
CM7	Volunteer moderators should be provided with materials for their roles.	User-to-user services that use volunteer moderation and are either: • Large, or • Multi-risk for content harmful to children.
SM3	Set and record internal content policies.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.
SM4	Set performance targets for search moderation functions.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.
SM5	Develop and apply policies on prioritisation of content for review.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.
SM6	Ensure search moderation functions are sufficiently resourced.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.
SM7	Ensure people working on search moderation receive training and materials.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.
US6	Provide age-appropriate user support materials for children.	User-to-user and search services that are multi-risk for content harmful to children.

23.37 If services are multi-risk (i.e., they have medium or high risk for two or more kinds of content harmful to children), we propose more demanding measures. These measures apply regardless of size, though we consider that large services are generally more likely to be multi-risk than smaller services.[1054]

23.38 Measures recommended for multi-risk services are intended to help mitigate risk in relation to all kinds of content harmful to children. They primarily consist of additional governance and content moderation measures[1055] involving more sophisticated processes, which we consider appropriate to manage and mitigate risk effectively in the more complex environment of a multi-risk service. Similar measures are recommended for U2U and search services. A further User Support measure (US6) recommends the provision of age-appropriate support materials. We believe this can help children better understand how to use a suite of different tools to reduce their exposure to different potential harms. However,

[1054] Large services typically have a higher number of users who are children than smaller services. As explained in Section 12, Volume 4 on Service risk assessment guidance and risk profiles, as part of assessing risk we recommend that services have regard to the number of children potentially affected by harmful content on the service. We indicate that a higher number of children on a service would tend to increase the potential impact from a given kind of harmful content, increasing the scope for a service to be medium to high risk of that content.

[1055] Note that the proposed reporting and complaints measures UR4(b) and UR5(b) include additional steps recommended for multi-risk services in relation to appeals, which are similar to some steps recommended under content moderation measures CM3 and CM4.

the benefit is likely to be small for users of services that pose risk of only one kind on content harmful to children and that typically have a more limited set of simple user tools.

23.39 We believe the measures allow for a reasonable degree of flexibility for services to determine a proportionate approach, meaning that the costs should vary depending on the characteristics of each service, with lower costs in general for smaller services with fewer risks. For example, costs associated with a well-resourced content moderation function are expected to scale with the size of a service, the volume of content, the number of risks and the level of risk (medium or high). This means that costs for a smaller service with medium risk for only two kinds of content harmful to children would be lower, compared to a service with high risk for many kinds of content. Similarly, the cost of tracking evidence of new or increasing harm is likely to depend on the size and complexity of a service, and the number of relevant harm vectors to be monitored. While we cannot quantify these costs precisely, our analysis indicates that costs can be expected to scale with the benefit of the measures across different services.

23.40 Nevertheless, we recognise that the combined costs for small and micro businesses that operate a multi-risk service can be considerable, particularly the costs of ensuring a well-resourced content moderation function, tracking its performance, and tracking evidence of new or increasing harm. Costs will be larger where services have few or none of the existing measures already in place, though this is also where a larger benefit would be expected due to the lack of existing safety measures. Certain costs will be somewhat reduced where businesses have already put in place similar measures proposed in our Illegal Harms consultation, but we consider that the incremental costs of our proposed draft Children's Safety Codes measures may still be substantial.

23.41 As noted in the previous sub-section, we recognise it is possible that some small and micro businesses may struggle to manage the combined costs of these measures and the ones discussed in earlier sections, which might result in some degradation in service quality or user experience. In extreme cases it is even possible that some services may withdraw from the UK. Other smaller services may choose to implement highly effective age assurance to stop children from accessing the service, which would avoid their having to implement any further measures for children's online safety.[1056] It is therefore possible that both children and adults in the UK may no longer be able to access some services. The flexibility we provide in the measures mitigates these risks of adverse effects to some degree when applying to multi-risk services.

23.42 On the other hand, we believe there is likely to be less benefit from extending these measures to smaller services that are low-risk or single-risk (i.e., they have a medium or high risk for a single kind of content harmful to children). These services operate in a simpler risk environment and could reasonably be expected to meet their child safety duties without employing more sophisticated formal processes and frameworks. Where these services are operated by small or micro businesses with relatively limited resources, children may benefit more from resources being channelled toward core activities such as moderating content, rather than diverted towards additional, more complex systems and processes that may have only small incremental benefits on such services. In any case, services that are not

[1056] Where a service implements highly effective age assurance, it can carry out a new children's access assessment to determine whether it is out of scope. Services not likely to be accessed by children must still comply with the duties about children's access assessments, which include a requirement to carry out a children's access assessment every year and sometimes more frequently. See Volume 2.

multi-risk remain in scope of core measures including user reporting, content moderation and governance measures (as outlined in 'Proposed measures recommended for all services'), but we allow them to implement these and meet their duties in a proportionate way.

23.43 In summary, while the cumulative cost of the proposed measures for smaller multi-risk services could be significant, our provisional view is that it would be proportionate, even taking into account the other measures they may be in scope of, as covered in previous sub-sections. We expect that these measures, when added to the baseline measures applied to all services, would be effective in further reducing harm to children on services that pose significant risks to them.

Impact of proposed measures for large services

23.44 All large services would be in scope of measures that apply to all services, as well as other measures discussed in the previous section, if they are multi-risk or meet the other specific criteria. However, some measures are recommended for services with specific reference to their size. This section considers the impact on large services by looking at the following categories of measures in turn:

a) Measures recommended for large services regardless of risk. (Applies to U2U and Search)
b) Measures recommended for large multi-risk services. (Applies to U2U and Search)

Additional proposed measures recommended for large services regardless of risk

23.45 The measures below are recommended for large services, regardless of their risk assessment. All large services, including those that are low-risk, would be in scope of these measures in addition to the core measures for all services (discussed previously under 'Proposed measures recommended for all services'). These additional measures consist primarily of extra governance and content moderation measures.

Table 23.4: Summary of additional proposed measures recommended for large services regardless of risk

No.	Description of proposed measure	Services we propose this will apply to
GA1	Most senior body to carry out and record an annual review of risk management activities relating to children's safety.	Large user-to-user services and large search services.
GA3	Written statements of responsibility for senior members who make decisions relating to management of child safety risks.	Search and user-to-user services that are either: • multi-risk for content harmful to children; or • large user-to-user services; or • large general search services.
GA5	Track unusual increases or new kinds of Primary Priority Content, Priority Content, and Non-designated Content on service.	Search and user-to-user services that are either: • multi-risk for content harmful to children; or • large user-to-user services; or • large general search services.

No.	Description of proposed measure	Services we propose this will apply to
GA6	Have a Code of Conduct that sets standards for employees around protecting children.	Search and user-to-user services that are either: • multi-risk for content harmful to children; or • large user-to-user services; or • large general search services.
GA7	Ensure staff involved in the design and operational management of service are sufficiently trained in approach to compliance with children's safety duties.	Search and user-to-user services that are either: • multi-risk for content harmful to children; or • large user-to-user services; or • large general search services.
CM2	Set internal content policies.	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.
CM3	Set performance targets for content moderation function.	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.
CM4	Have and apply policies on prioritisation of content for review.	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.
CM5	Ensure content moderation functions are well-resourced.	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.
CM6	Ensure content moderation teams are appropriately trained	User-to-user services that are: • Large, or • Multi-risk for content harmful to children.
CM7	Volunteer moderators should be provided with materials for their roles.	User-to-user services that use volunteer moderation and are either: • Large, or • Multi-risk for content harmful to children.
SM2	When a user is believed to be a child, filter identified Primary Priority Content out of their search results through a safe search setting. Users believed to be a child should not be able to turn this setting off.	Large general search services.
SM3	Set and record internal content policies.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.
SM4	Set performance targets for search moderation functions.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.
SM5	Develop and apply policies on prioritisation of content for review.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.
SM6	Ensure search moderation functions are sufficiently resourced.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.
SM7	Ensure people working on search moderation receive training and materials.	Search services that are: • Large general search services, or • Multi-risk for content harmful to children.

No.	Description of proposed measure	Services we propose this will apply to
SD2	When a user is believed to be a child, filter identified Primary Priority Content out of their search results through a safe search setting. Users believed to be a child should not be able to turn this setting off.	Large general search services.

23.46 For each of these measures, we have explained in the relevant section of the consultation why we propose recommending these for large services, even if they assess as low-risk. We expect that large services will tend to be relatively complex and multi-faceted, and to have large volumes of content. For such services to protect children, we consider it proportionate that they should take additional steps that promote effective governance and content moderation, since a failure to do so may lead to many children experiencing harm. As the nature of risks and kinds of content harmful to children can change over time, having suitable governance and content moderation measures[1057] in place can help manage new and escalating risks quickly and effectively.

23.47 For many of these proposed measures, the costs of implementing them are likely to be lower if a service is low risk. We have also taken into account that many services will also be in scope of similar measures in the Illegal Harms Codes. This may reduce the cost of implementation – for example, where existing governance frameworks or content moderation processes may be used or adapted to implement the measures in the draft Children's Safety Codes. We therefore expect the measures to be proportionate even for large, low-risk services.

User-to-user services

23.48 The additional governance measures recommended for large services are considered proportionate, even for large, low-risk services, because the consequences of a governance failure in a service with a large user base can have a particularly significant impact. Although the potential cost associated with some of the governance measures can be substantial, we consider the cost to be proportionate given the resources that large services are likely to have. For services that are also in scope of the Illegal Harms Codes, we believe that some of the governance measures will overlap with the related Illegal Harms measures, meaning services will only face an incremental cost to extend these measures as recommended for protection of children.

23.49 Large U2U services will also face additional content moderation measures. We consider these to be proportionate for large services as they are likely to have larger volumes of content and reports. In the absence of more sophisticated moderation resources and processes, there may be a higher likelihood of moderation failures that could lead to many children experiencing harm on a service. These measures are also important for large services to have an adequate understanding of their risk environment. These additional measures are still defined flexibly and allow for suitable approaches based on a service's specific needs and context. Many of the relevant costs would be lower for a large, low-risk service than for a large service which poses several significant risks to children. The latter

[1057] Note that the proposed reporting and complaints measures UR4(b) and UR5(b) include additional steps recommended for large services in relation to appeals, which are similar to some steps recommended under content moderation measures CM3 and CM4.

service would be expected to require additional resources and more extensive training, for example, to implement the measures.

Search services

23.50 Large general search services will also need to implement the equivalent governance and content moderation measures to those described above for U2U services for similar reasons.

23.51 There are additional measures that apply to large general search services. These recommend steps to implement immutable safe search settings that filter PPC for users believed to be a child, provide crisis prevention information to users, and take action in response to predictive search suggestions that present a risk of children encountering PPC and PC. We believe these to be proportionate without reference to the outcome of these services' risk assessment. We consider these measures to be important in protecting children, given that such services are likely to be inherently risky for children given their wide user reach. While their cost is likely to be substantial, we expect that large general search services have sufficient capacity to implement the measures.

23.52 We propose to not recommend these measures for vertical search services. We believe such services are inherently less likely to present risks of harm to children than general search services, while they typically have greater control over content shown to users than general search services. Any benefits of recommending such measures to these services would therefore be low and we do not consider it proportionate.

23.53 Note that there is an additional measure (TS3), recommending that service providers should summarise the findings of their most recent children's risk assessment in their terms or statements. This measure reflects a specific requirement in the Act for Category 1 and 2A services, over which we have not exercised any material discretion. We do not include this measure in the tables in this section, as its applicability to different kinds of services will depend on future secondary legislation to define thresholds for Category 1 and 2A services. Our categorisation advice to the Secretary of State proposed that Category 1 and 2A services should, among other criteria, have a number of UK users that is consistent with those services considered large services.[1058]

Additional proposed measures recommended for large and risky services

23.54 The measures below are recommended for services which are large and which also meet relevant risk criteria. These are additional to the core measures recommended for all services (discussed under 'Proposed core measures recommended for all services'), the targeted measures recommended for services of any size that meet relevant risk criteria (discussed under 'Additional proposed measures recommended for services that meet specific risk criteria') and the measures recommended for all large services, regardless of risk (discussed in the previous sub-section).

Table 23.5: Summary of additional proposed measures recommended for large and risky services

No.	Description of proposed measure	Services we propose this will apply to
GA4	Have an internal monitoring and assurance function to provide independent assurance that measures are effective.	User-to-user services that are: • Large, and • Multi-risk for content harmful to children.

[1058] Ofcom, March 2024, Categorisation: Advice submitted to the Secretary of State.

No.	Description of proposed measure	Services we propose this will apply to
RS3	Enable children to provide negative feedback on content that is recommended to them.	User-to-user services that: • Operate a content recommender system, and • Are medium or high risk for at least two kinds of Primary Priority Content and/or Priority Content (excluding bullying),[1059] and • Are large.
US4	The provision of information to child users when they restrict interactions with other accounts or content.	User-to-user services that: • Have certain functionalities that restrict action against another account or content[1060] • Are large, and • Are multi-risk for content harmful to children.
US5	Signpost children to support at key points in the user journey.	**Intervention point 2 – when children post or re-post content** Large user-to-user services that: • Have posting/re-posting functionalities, and • Are medium or high risk of one or more of: suicide content, self-harm content, eating disorder content, or bullying content, and • Have measures that enable them to identify when a user posts or re-posts suicide, self-harm, eating disorder or bullying content.

23.55 We propose an additional cross-cutting governance measure for services that are large and multi-risk, recommending an internal monitoring and assurance function to independently assess the effectiveness of the mitigations of content harmful to children. This is likely to be a costly measure which would not be proportionate if applied to smaller and less complex multi-risk services that are already in scope of the other governance measures mentioned previously, and for whom the need for this function would be more limited. However, for large multi-risk services, we consider that adding this measure on top of other governance measures is proportionate. The benefits, in terms of supporting effective protection of children across a potentially complex service with a large user base, are likely to be greater, while such services are likely to be able to access necessary resources to implement the measures.

23.56 There are further measures that only apply to large U2U services where these services have specific functionalities or risks, including an additional Recommender Systems measure to provide children with a means to express negative sentiment towards content and have this feed into recommender systems. This measure would involve substantial costs if also recommended for smaller services alongside the Recommender System measures and other recommended measures for smaller services. We also consider that smaller services may lack the capacity to implement the measure as effectively as larger services, such that the

[1059] We are also minded to extend this measure for two potential kinds of Non-designated Content. See Section 20 on Recommender Systems for more information.
[1060] Please refer to Section 21, Volume 5 on User Support for more information on the functionalities that are applicable to Measure US4.

benefits of the measure in terms of protecting children could be significantly lower on smaller services.

23.57 There are further measures for large services related to providing information to users. We recommend that large services with relevant risks should signpost users to support resources when they post, share or search for certain content. Separately, we recommend the provision of supportive information when users take action against another user or a piece of content. While we consider these measures have the potential to improve children's safety online, their costs are uncertain, and we expect they may be material for smaller services. On balance, we consider that these measures could have a lesser effect on children's online safety on smaller services, given the range of measures already applicable to smaller, risky services. We have prioritised measures for smaller services that we believe can deliver material improvement in children's safety. We therefore provisionally conclude that the measures described in this sub-section for large services would be disproportionate for smaller services.

24. Statutory tests

In designing our codes of practice, the Online Safety Act 2023 ('the Act') requires us to have regard to principles, online safety objectives and content requirements of the codes of practice, set out in Schedule 4 to the Act. The Communications Act 2003 ('CA 2003') also sets out duties that we must fulfil when exercising our regulatory functions, including the online safety functions.

In this chapter we outline the section 3 duties of the CA 2003 that we must fulfil in carrying out our regulatory functions, the principles and objectives set out in Schedule 4 to the Act, and explain the reasons why our proposals, in particular our proposed recommendations for our draft Children's Safety Codes ('Codes'), meet these requirements. We provide further detail regarding Ofcom's duties relating to the preparation of our Codes in our Legal Framework (Annex 13).

Consultation questions

59. Do you agree that our proposals, in particular our proposed recommendations for the draft Children's Safety Codes, are appropriate in the light of the matters to which we must have regard? If not, please explain why.

Background

24.1 The CA 2003 places a number of duties on us in carrying out our functions, including requiring us to have regard to the risk of harm to citizens presented by content on regulated services and the need for a higher level of protection for children than for adults. Further, in designing our draft Codes, the Act requires us to have regard to a number of principles and objectives, and content requirements of the Children's safety codes set out in Schedule 4 to the Act.

24.2 In Sections 15-22, we set out our proposed recommendations; an overview of these recommendations can be found in Section 13, and our combined assessment of the proposed measures can be found in Section 23. The draft Children's Safety Codes themselves can be found in full in Annex 7 (U2U) and Annex 8 (search).

24.3 We consider that our proposals meet the requirements set out in section 3 of the CA 2003 and Schedule 4 to the Act. In this section, we take each of the requirements in turn and set out how we have met them in reaching our set of proposed recommendations.

Duties and principles

The Communications Act 2003

23.58 The Communications Act 2003 places a number of duties on us that we must fulfil when exercising our regulatory functions, including our online safety functions. As required by section 3 of the CA 2003, in making the proposals in this consultation, including the proposed recommendations in our draft Codes, we have had regard to the matters set out below.

Section 3(1): It shall be the principal duty of Ofcom, in carrying out their functions: a) to further the interests of citizens in relation to communication matters; and b) to further the interests of consumers in relevant markets, where appropriate by promoting competition.

- We have clearly identified how proposed measures for the draft Children's Safety Codes will mitigate risks of harm to children online, thereby furthering their interests, as well as the interests of citizens in the UK more generally. Much of what we know about the risk of harm to children comes from engaging with children. As part of our research, children told us what they want and need to ensure they can live a safer life online, including the measures they would like to see service providers implement.
- We have considered the interests of consumers in relevant markets (particularly users of regulated services) as part of our assessment of the proportionality of our proposals, including any potential impacts on the provision of services to users.

Section 3(3): In performing their duties under subsection (1), Ofcom must have regard in all cases to (a) to the principles under which regulatory activities should be transparent, accountable, proportionate, consistent and targeted only at cases in which actions is needed, and (b) any other principles appearing to us to represent best regulatory practice.

- In the interest of transparency, accountability and fairness (and as required under the Act), we are consulting stakeholders on our proposals and publishing impact assessments for each of the measures we are proposing to include in the Children's Safety Codes. We are setting out clearly the evidence and assumptions used to arrive at our proposals. We have also conducted an impact assessment for our draft Children's Risk Assessment Guidance.
- Our impact assessments of proposed measures consider effectiveness, costs, rights, and other relevant factors and explain why we consider the proposed measures are proportionate. We consider the proportionality of the package of our measures as a whole in our combined impact assessment in Section 23. Our impact assessment for the draft Children's Risk Assessment Guidance is set out in Section 12. See our impact assessment guidance for more information on how we approach impact assessments.
- Our proposed measures are informed by our assessment of the risks of harm to children (Volume 3). We have prioritised developing proposed measures that can effectively mitigate the significant risks identified in our analysis and those required by the Act and have targeted our proposed measures at the kinds of services which we think should be deploying them because this would lead to the greatest benefits given the risks they pose. Similarly, we consider that the proposed approach set out in our draft Children's Risk Assessment Guidance is a proportionate approach to ensuring that services understand the risks that they pose to children.

Section 3(2)(g): In carrying out our functions, Ofcom are required to secure (g) the adequate protection of citizens from harm presented by content on regulated services, through the appropriate use by providers of such services of systems and processes designed to reduce the risk of such harm.

- Our proposals set out steps we consider service providers should take to assess and mitigate risks that content on their services pose to children. They are

informed by our own assessment of the risks of harm to children. Our proposed measures are designed to reduce the risk of harm to children from content harmful to children, namely Primary Priority Content that is harmful to children ('PPC'), Priority Content that is harmful to children ('PC') and Non designated content ('NDC').

- Proposed measures in relation to governance and accountability (Section 11), content moderation (Section 16), search moderation (Section 17), user reporting and complaints (Section 18) specifically concern the safer design and functioning of processes, and Terms of Service and Publicly Available Statements (Section 19).
- Proposed measures in relation to age assurance (Section 15), recommender systems (Section 20), user support (Section 21) and search features, functionalities and user support (Section 22) concern safer systems and functionalities on services.

24.4 In relation to matters to which section 3(2)(g) in the CA 2003 is relevant, Section 3(4A) sets out that in performing their duties under subsection (1), Ofcom must have regard to such of the following as appear to them to be relevant in the circumstances:

(a) The risk of harm to citizens presented by content on regulated services.

- Our Children's Register of Risks (Section 7) sets out the risks of harm to children from content harmful to children as we assess it to manifest in the current environment. These risks, alongside findings from services' children's risk assessments, largely inform what proposed measures will be appropriate for a service provider to address the risk of harm to children.

(b) The need for a higher level of protection for children than for adults.

- All the proposals in our consultation are designed to achieve this outcome. In particular, our proposed measures are intended to enable regulated services to effectively manage and mitigate the risk of harm to children from content harmful to children, which does not apply in the same way or may not have the same impacts on adults.
- Proposed measures for recommender systems (filtering out content likely to be PPC and limiting the prominence of content likely to be PC) are recommended only for users who are children. This is enabled through some of our proposed age assurance measures (Section 15), which will facilitate a means for services to distinguish adults from children. Our proposed age assurance measures will also allow for other forms of access and content controls to be targeted at children for relevant services.
- Proposed measures that are not targeted at children specifically, are still framed in ways to improve the experience of children. For example, proposed measures under User reporting and complaints will mean services will have to accept reports of content harmful to children from all users, but at the same time our proposals will ensure that reports are processed and managed in a way that will break down barriers to reporting that we have identified children specifically face.
- Proposed measures in the draft Children's Safety Codes that build on proposals in the draft Illegal Content Codes go further to address risks faced by children. We are proposing changes for the Illegal Content Codes to also increase protections for children from illegal content (see user reporting and complaints Section 18,

content moderation Section 16, terms of service and publicly available statements Section 19).

(c) The need for it to be clear to providers of regulated services how they may comply with their duties set out under the Act.

- Our proposals aim to provide clarity and tangible steps that services can take to meet their duties in the Act. We have clearly explained that the Act provides that services likely to be accessed by children and which choose to implement the measures we recommend in the draft Children's Safety Codes, will be considered as complying with relevant duties. Our draft Children's Risk Assessment Guidance is intended to help services understand how they can comply with their duties to carry out a children's risk assessment.

(d) The need to exercise their functions so as to secure that providers of regulated services may comply with such duties by taking measures, or using measures, systems or processes, which are (where relevant) proportionate to (i) the size or capacity of the provider in question, and (ii) the level of risk of harm presented by the service in question, and the severity of the potential harm.

- We have clearly identified in our draft Codes which measures apply to what types and sizes of services, for the reasons given in each relevant section of this consultation, with more demanding expectations placed on services that pose greater risk of harm to children, even if they are smaller services. We also propose a minority of our measures for large services only, because we do not consider they would be proportionate to be applied to smaller services.
- Where appropriate our measures are designed to give a degree of flexibility so that services can tailor their approach to their context, taking into account factors including their size and capacity.
- Similarly, our draft Children's Risk Assessment Guidance proposes an approach to the children's risk assessment which takes account of the nature and size of services, for example in deciding what evidence to take into consideration.

(e) & (f) The desirability of promoting the use by providers of regulated services of technologies which are designed to reduce the risk of harm to citizens presented by content on regulated services; and the extent to which providers demonstrate, in a way that is transparent and accountable, that they are complying with their duties.

- Our proposals allow services flexibility to implement technologies in a way that is cost-effective and proportionate to the circumstances of the service. For example, our proposed measures for content moderation and search moderation (see Section 16 and Section 17) stipulate services may use a combination of automated tools and human review to moderate content. We also provide flexibility to services in how they can implement proposed measures around age assurance and do not prescribe specific technologies. We provide draft guidance on highly effective age assurance at Annex 10, which includes the criteria any chosen technology would need to meet to be highly effective age assurance, together with draft Guidance on Content Harmful to Children.

24.5 Section 3(4) of the CA 2003[1061] sets out other matters to which Ofcom must, to the extent they appear to us relevant in the circumstances, have regard, in performing our duties.

Section 3(4) : Ofcom must also have regard, in performing those duties, to such of the following as appear to them to be relevant in the circumstances […] (b) the desirability of promoting competition in relevant markets, (d) the desirability of encouraging investment and innovation in relevant markets; (h) the vulnerability of children and of others whose circumstances appear to Ofcom to put them in need of special protection; (i) the needs of persons with disabilities, of the elderly and of those on low incomes; (j) the desirability of preventing crime and disorder; (k) the opinions of consumers in relevant markets and of members of the public generally; (l) and the different interests of persons in the different parts of the United Kingdom, of the different ethnic communities within the United Kingdom and of persons living in rural and urban areas.

- Where appropriate, in proposing measures, we have had regard to the desirability of promoting competition and encouraging investment and innovation. A number of our proposed measures accordingly provide flexibility for services to decide how to achieve compliance. As set out above, we have considered the interests of consumers in relevant markets as part of our impact assessments of proposed measures, including any indirect impacts on consumers in cases where our measures could affect competition, investment and innovation in respect of the online services that they use. In proposing measures and draft guidance, we have had regard to the objective of a higher standard of protection for children than for adults, assessing whether measures are expected to be effective at achieving this. Under our equality impact assessments across the proposed measures and draft guidance, we have considered the needs of persons of protected and listed characteristics. We have also considered our Welsh language obligations. See Annex 14.

Schedule 4, Online Safety Act 2023

24.6 As required by paragraph 1 of Schedule 4 to the Act, we have considered the appropriateness of applying provisions of the draft Children's Safety Codes to different kinds and sizes of Part 3 services and to providers of differing sizes and capacities and has set out in this consultation our reasons for proposing to apply some Codes recommendations to services of different kinds, sizes and capacities.

24.7 We have had regard to the following principles in Schedule 4, as follows:

Paragraph 2(a): providers of Part 3 services must be able to understand which provisions of the code of practice apply in relation to a particular service they provide.

- We have clearly identified in our draft Codes which measures apply to what types and sizes of services, for the reasons given in each relevant section of this consultation. In our summary of proposed codes measures we provide an overview at a glance of proposed measures and the services we propose they apply to.

Paragraph 2(b): the measures described in the code of practice must be sufficiently clear, and at a sufficiently detailed level, that providers understand what those measures entail in practice.

[1061] As amended by section 82 of the Act

- Having regard to the need for it to be clear to providers of regulated services how they may comply with their duties dealt with in this consultation, we have aimed to be as clear and detailed as possible in our draft Codes, consistent with acting proportionately.

Paragraph 2(c): the measures described in the code of practice must be proportionate and technically feasible: measures that are proportionate or technically feasible for providers of a certain size or capacity, or for services of a certain kind or size, may not be proportionate or technically feasible for providers of a different size or capacity or for services of a different kind or size;

- We have clearly identified in our draft Codes which measures apply to what types and sizes of services, for the reasons given in each relevant section of the Consultation. We have considered proportionality and technical feasibility, where appropriate, as part of our impact assessment across this consultation. This includes taking into account evidence of current practice by user-to-user and search service providers who are already taking steps that are similar or related to measures that we propose. We consider effectiveness, costs, rights impacts, and other relevant factors in our assessment of proportionality. The more demanding proposed measures, we recommend for services that pose greater risk of harm to children, even if they are smaller services. At the same time, certain measures are recommended for large services only, based on proportionality considerations including with respect to the capacity of smaller services to implement. For further detail on our approach for which measures we proposed to apply to what services, please see the framework for codes (Section 14).

Paragraph 2(d): the measures described in the code of practice that apply in relation to Part 3 services of various kinds and sizes must be proportionate to Ofcom's assessment under section 98 of the risk of harm presented by services of that kind or size.

- Our reasoning to support the proposed recommendations, identifies the relevant risks of harm that our measures address, and explains why we consider each proposed measure is proportionate in the light of those harms. As required by section 3(4A)(b)(ii) of the CA 2003, in considering proportionality we have had regard to the severity of the potential harm as well as the level of risk of harm, as identified in our draft Children's Register of Risks (Section 7). Where appropriate, we have clearly identified in our draft Codes which measures would apply to what types and sizes of services, for the reasons given in each relevant section of this consultation. Overall, our draft Codes place more demanding expectations on services that pose greater risk of harm to children, even if they are smaller services, because this is where measures have the greatest potential to support safer experiences for children online.

24.8 Having had regard to the desirability of promoting the use by providers of regulated services of technologies which are designed to reduce the risk of harm to citizens presented by content on regulated services, and to the seriousness of the harms concerned, our proposals do not recommend specific technologies at this time due to limited evidence. This allows services to act in accordance with our recommendations using any appropriate technology or input.

24.9 Having regard to the desirability of encouraging investment and innovation in the markets for regulated services and these technologies, our proposals provide sufficient flexibility for

services and by not recommending specific technologies or the use of specific inputs, in order to secure that services can act in accordance with our recommendations using any appropriate technology or input.

Ofcom's online safety objectives

U2U services

24.10 As required by paragraph 3 of Schedule 4 to the Act, we have also ensured that the proposed recommendations are compatible with the pursuit of the applicable online safety objectives for U2U services as follows:

Paragraph 4(a)(i): a service should be designed and operated in such a way that the systems and processes for regulatory compliance and risk management are effective and proportionate to the kind and size of service.

- In Section 11 (governance and accountability), we have set out the governance measures which we propose to recommend having regard, among other things, to the kind and size of service. We consider these to be compatible with this objective.

Paragraph 4(a)(ii): a service should be designed and operated in such a way that the systems and processes are appropriate to deal with the number of users of the service and its user base.

- As set out in our overview, we have considered the size of services in our assessment of whether the recommendation of certain measures is proportionate; in Section 11 (governance and accountability), Section 16 (content moderation), Section 18 (User reporting and complaints), and Section 21 (user support), we have set out the systems and processes measures which we propose to recommend having regard, among other things, to the number of users of the service and its user base. We consider these to be compatible with this objective.

Paragraph 4(a)(iii): a service should be designed and operated in such a way that United Kingdom users (including children) are made aware of, and can understand, the terms of service.

- In Section 19 (terms of service and publicly available statements) we are consulting on proposed recommendations which we consider would be compatible with this objective, namely for terms regarding the protection of children to contain all information mandated by the Act (Measure TS1) as well as to be clear and accessible (Measure TS2).

Paragraph 4(a)(iv): a service should be designed and operated in such a way that there are adequate systems and processes to support United Kingdom users.

- In Section 18 (user reporting and complaints), Section 20 (recommender systems) and Section 21 (user support), we are consulting on proposed recommendations which we consider would be compatible with this objective.

Paragraph 4(a)(vi): a service should be designed and operated in such a way that the service provides a higher standard of protection for children than for adults.

- Having regard to the need for a higher standard of protection for children than for adults, we consider our proposed recommendations would be compatible with this objective.

Paragraph 4(a)(vii): a service should be designed and operated in such a way that the different needs of children at different ages are taken into account.

- In Section 21 (user support) we set out some considerations in respect of the needs of children at different ages. However, as per our draft Children's Register of Risks, more evidence is needed to understand in greater detail the risks of harm that children in different age groups face, to be able to provide robust recommendations for measures to address those bespoke risks. We discuss this further in the Introduction to Volume 3 at Section 7.

Paragraph 4(a)(viii): a service should be designed and operated in such a way that there are adequate controls over access to the service by adults.

- We are not proposing additional measures that address this objective given that this consultation focuses on providing specific protections to children; therefore, we have focused on controls on access by children where appropriate and proportionate to protect them from harm (see Section 15 on age assurance). We have considered access controls for adult users in our Illegal Harms Consultation. In Chapter 21 (User Access) in our Illegal Harms Consultation we set out why we do not consider it appropriate to restrict access to services generally by adults. We explained the measures we proposed to limit the activities of proscribed organisations. In Chapter 20 (Enhanced User Controls) in our Illegal Harms Consultation we proposed a measure setting out the steps we recommend a service to take if it purports to offer a verification scheme for users.

Paragraph 4(a)(ix): a service should be designed and operated in such a way that there are adequate controls over access to, and use of, the service by children, taking into account use of the service by, and impact on, children in different age groups.

- In Section 15 (age assurance) we are consulting on proposed recommendations which we consider would be compatible with this objective and explain how we have taken into account use of the service by, and impact on, children in different age groups. For more information on children in different age groups and risks see the Children's Register of Risks at Section 7.

Paragraph 4(b): a service should be designed and operated so as to protect individuals in the United Kingdom who are users of the service from harm, including with regard to:

- algorithms used by the service,
- functionalities of the service, and
- other features relating to the operation of the service.

- All our recommendations seek to protect users, specifically children, from harm. In particular, in Section 11 (governance and accountability), Section 15 (age assurance), Section 16 (content moderation), Section 18 (user reporting and complaints), Section 21 (user support), and Section 20 (recommender systems),

we are consulting on proposed recommendations which we consider would be compatible with this objective.

24.11 We are not at this stage consulting on measures relating to paragraph 4(a)(v) – "(in the case of a Category 1 service) users are offered options to increase their control over the content they encounter and the users they interact with" - given it is specific to Category 1 services only. We will explore proposed measures for categorised services in greater detail in Phase 3 of Ofcom's work.

Schedule 4 requirements on Content of Codes of Practice: age assurance

24.12 Schedule 4 paragraph 12(1) to the Act states that the paragraph is about the inclusion of age assurance in a code of practice as a measure recommended for the purpose of compliance with any of the duties set out in section 12(2) or (3) or 29(2) or (3), and sub-paragraph (2) sets out some further principles in addition to those in paragraphs 1 and 2 (general principles) and 10(2) (freedom of expression and privacy), which are particularly relevant.

Paragraph 12(3): In deciding whether to recommend the use of age assurance, or which kinds of age assurance to recommend we must have regard to the following:

(a) The principle that age assurance should be effective at correctly identifying the age or age-range of users;

(b) The relevant standards set out in the latest version of the code of practice under section 123 of the Data Protection Act 2018 (age-appropriate design code);

(c) The need to strike the right balance between (i) the level of risk and the nature, and severity, of potential harm to children which the age assurance is designed to guard against, and (ii) protecting the right of users and (in the case of search services or the search engine of combined services) interested persons to freedom of expression within the law;

(d) The principle that more effective kinds of age assurance should be used to deal with higher levels of risk of harm to children;

(e) The principle that age assurance should be easy to use, including by children of different ages and with different needs;

(f) The principle that age assurance should work effectively for all users regardless of their characteristics or whether they are members of a certain group

(g) The principle of interoperability between different kinds of age assurance.

- In Section 15, we discuss our proposed measures regarding age assurance, including how we have had regard to factors (a) – (g) above in developing the policy that has informed the proposed measures and draft guidance.

Search services

24.13 As required by paragraph 3 of Schedule 4 to the Act, we have ensured that the proposed recommendations are compatible with the pursuit of the applicable online safety objectives for search services as follows:

Paragraph 5(a)(i): a service should be designed and operated in such a way that the systems and processes for regulatory compliance and risk management are effective and proportionate to the kind and size of service.

- In Section 11 (governance and accountability), we have set out the governance measures which we propose to recommend having regard, amongst other things, to the kind and size of service. We consider these to be compatible with this objective.

Paragraph 5(a)(ii): a service should be designed and operated in such a way that the systems and processes are appropriate to deal with the number of users of the service and its user base.

- In Section 11 (governance and accountability), Section 17 (search moderation), and Section 22 (search features, functionalities and user support), we have set out the systems and processes measures which we propose to recommend having regard, among other things, to the number of users of the service and its user base. We consider these to be compatible with this objective.

Paragraph 5(a)(iii): a service should be designed and operated in such a way that United Kingdom users (including children) are made aware of, and can understand, the publicly available statement referred to in sections 23 and 25.

- In Section 19 (terms of service and publicly available statements) we are consulting on a proposed recommendations which we consider would be compatible with this objective. In making these recommendations, our duty to have regard to the extent to which providers of regulated services demonstrate, in a way that is transparent and accountable, that they are complying with their duties set out in the Act, is relevant.

Paragraph 5(a)(iv): a service should be designed and operated in such a way that there are adequate systems and processes to support United Kingdom users.

- In Section 17 (search moderation), and Section 22 (search features, functionalities and user support) we are consulting on proposed recommendations which we consider would be compatible with this objective.

Paragraph 5(a)(v): a service should be designed and operated in such a way that the service provides a higher standard of protection for children than for adults.

- Having had careful regard to the need for a higher level of protection for children than for adults, in Section 17 (search moderation) and Section 22 (search features, functionalities and user support) we are consulting on proposed recommendations which we consider would be compatible with this objective.

Paragraph 5(a)(vi): a service should be designed and operated in such a way that the different needs of children at different ages are taken into account.

- In Section 21 (user support) – specifically Measure US6 - we set out some considerations in respect of the needs of children at different ages. However, as per our Children's Register of Risks, more evidence is needed to understand in greater detail the risks of harm that children in different age groups face, to be

able to provide robust recommendations for measures to address those bespoke risks. We discuss this further in the Introduction to Volume 3 at Section 7.

Paragraph 5(b): a service should be assessed to understand its use by, and impact on, children in different age groups.

- Service providers have a duty to assess their user base, including the number of children in different age groups on the service. Additionally, service providers must assess the impact of the risk of harm to children in different age groups on their services – see Children's Risk Assessment Guidance at Section 12 in Volume 4. The Children's Register of Risks and Children's Risk Profiles include further guidance on the developmental stages of children in different age groups in the context of content harmful to children, to help services consider the risk of harm to children (see Section 6).

Paragraph 5(c): a search engine should be designed and operated so as to protect individuals in the United Kingdom who are users of the service from harm, including with regard to:

- algorithms used by the search engine,
- functionalities relating to searches (such as a predictive search functionality), and
- the indexing, organisation and presentation of search results

- In Section 11 (governance and accountability), Section 17 (search moderation) and Section 22 (search features, functionalities, and user support) we are consulting on proposed recommendations which we consider would be compatible with this objective.

Content of codes of practice

U2U services

24.14 Codes of practice that describe measures recommended for the purpose of compliance with a duty set out in section 12(2) or (3) of the Act (children's online safety) must include measures in each of the areas of a service listed in section 12(8). This provision applies to the extent that inclusion of the measures in question is consistent with:

a) Ofcom's duty to consider the appropriateness of provisions of the code of practice to different kinds and sizes of Part 3 services and to providers of differing sizes and capacities;

b) The principle that the measures described in the code of practice must be proportionate and technically feasible: measures that are proportionate or technically feasible for providers of a certain size or capacity, or for services of a certain kind or size, may not be proportionate or technically feasible for providers of a different size or capacity or for services of a different kind or size; and

c) the principle that the measures described in the code of practice that apply in relation to Part 3 services of various kinds and sizes must be proportionate to Ofcom's assessment (under section 98) of the risk of harm presented by services of that kind or size.

24.15 We have made proposals for U2U services in each of the areas of a service listed in section 12(8) as follows:

a) regulatory compliance and risk management arrangements – see Section 11 (governance and accountability)

b) design of functionalities, algorithms and other features – see Section 18 (user reporting and complaints), Section 20 (recommender systems) and Section 21 (user support).

c) policies on terms of use – see Section 19 (terms of service and publicly available statements), and Section 16 (content moderation)

d) policies on user access to the service or to particular content present on the service, including blocking users from accessing the service or particular content – see Section 15 (age assurance)

e) content moderation, including taking down content – see Section 16 (content moderation)

f) functionalities allowing users to control the content they encounter – see Section 21 (user support) and Section 20 (recommender systems – specifically proposed Measure RS3)

g) user support measures – see Section 21 (user support)

h) staff policies and practices – see Section 11 (governance and accountability), and Section 16 (content moderation)

24.16 Proposed measures have been assessed for their impact on users' rights in line with paragraph 10(1)-(3) of the Act which requires measures described in a code of practice which are recommended for the purpose of compliance with any of the relevant duties, to be designed in the light of the following principles:

a) The importance of protecting the rights of users and (in the case of search services or combined services) interested persons to freedom of expression within the law.
b) The importance of protecting the privacy of users.

24.17 All the measures we propose for the draft Children's Safety Codes, in line with paragraph 11, relate only to the design or operation of a Part 3 service (a) in the United Kingdom, or (b) as it affects United Kingdom users of the service.

Search services

24.18 Codes of practice that describe measures recommended for the purpose of compliance with a duty set out in section 29(2) or (3) of the Act (children's online safety) must include measures in each of the areas of a service listed in section 29(4). This provision applies to the extent that inclusion of the measures in question is consistent with:

a) Ofcom's duty to consider the appropriateness of provisions of the code of practice to different kinds and sizes of Part 3 services and to providers of differing sizes and capacities;

b) the principle that the measures described in the code of practice must be proportionate and technically feasible; and

c) the principle that the measures described in the code of practice that apply in relation to Part 3 services of various kinds and sizes must be proportionate to Ofcom's

assessment (under [section 89]) of the risk of harm presented by services of that kind or size.

24.19 We have made proposals for search services in the following areas of a service listed in section 29(4) as follows:

a) regulatory compliance and risk management arrangements – see Section 11 (governance and accountability)
b) design of functionalities, algorithms and other features relating to the search engine – see Section 18 (user reporting and complaints), Section 22 *(*search features, functionalities and user support)
c) user support measures – see Section 21 (user support) specifically Measure US6, and Section 22 (search features, functionalities and user support)
d) staff policies and practices – see Section 11 (governance and accountability), and Section 17 (search moderation).
e) Functionalities allowing users to control the content they encounter in search results – see Section 17 (search moderation), specifically the proposed 'safe search' measure (Measure SM2).

24.20 Proposed measures have been assessed for their impact on users' rights in line with paragraph 10(1)-(3) of Schedule 4 to the Act which requires measures described in a code of practice which are recommended for the purpose of compliance with any of the relevant duties, to be designed in the light of the following principles:

f) The importance of protecting the rights of users and (in the case of search services or combined services) interested persons to freedom of expression within the law.
g) The importance of protecting the privacy of users.

24.21 All the measures we propose for the draft Children's Safety Codes, in line with paragraph 11, relate only to the design or operation of a Part 3 service (a) in the United Kingdom, or (b) as it affects United Kingdom users of the service.