Additional Safety Measures

Online Safety

Consultation

Published 30 June 2025
Closing date for responses: 20 October 2025

For more information on this publication, please visit ofcom.org.uk

Contents

Section

1. Overview

2025 is Ofcom's 'Year of Action' for online safety. Our Illegal Content Codes of Practice are already in force, and our Protection of Children Codes come into effect in July. Service providers are required to complete their risk assessments setting out how they will ensure users are kept safer, and we have launched enforcement action against platforms where we have reason to believe they may not be compliant with their new duties.

We are now setting out proposals for a series of additional safety measures to further strengthen our 'first-edition' Codes. They include wider use of automated technologies to detect harm and additional steps to ensure that services are safer by design. Our proposals take account of the latest developments in harms and technologies, as well as a range of evidence captured through our ongoing engagement with civil society and other stakeholders.

Stopping illegal content going viral

We want to ensure platforms are doing more, earlier, to prevent serious illegal content from spreading. That means having protocols in place to respond to spikes in illegal content during a crisis, like the riots we saw after the Southport attacks last year. It means users should not be recommended potentially illegal material until it has been checked by services. It means better protections around livestreams, with better reporting functions and human moderators available at all times. And it means taking action against users who share or upload illegal content and content harmful to children.

Tackling harms at source

Huge volumes of content appear online every day, and providers need to make better use of technology to prevent illegal material from reaching users. We are proposing they use a technology called hash matching to identify intimate image abuse and terrorism content. We also think some services should go further – assessing the role that automated tools can play in detecting a wider range of content, including child abuse material, fraudulent content and content promoting suicide and self-harm and implementing new technology where it is available and effective.

Affording further protections to children

We know children remain at risk of some of the most egregious harms online. That is why we are proposing increased highly effective age assurance to help protect children from grooming. We are acting on risky environments like livestreaming, recommending that users should no longer be able to interact with children's livestreams through comments and gifts. And we are proposing that platforms take action to prevent individuals who share child sexual abuse material from using the service.

We are now seeking feedback on these proposals by 20 October 2025. Whether you are an online service provider, parent, policymaker, civil society organisation, or someone with lived experience, all responses will be carefully considered to inform our final decisions.

2. Summary of Our Proposals

The table below sets out a summary of the proposed Code measures within this consultation.

Proposed measures	Who should implement this	Relevant Code(s)
Livestreaming		
ICU D17: The provider should have a mechanism to enable users to report that a livestream contains content that depicts the risk of imminent physical harm.	All user-to-user services offering one-to-many livestreaming at medium or high risk of terrorism content, grooming, image-based CSAM, assisting or encouraging suicide, hate, or harassment, stalking, threats and abuse offences.	Illegal Content User-to-User Codes
ICU C16: The provider should, as part of its content moderation function (see ICU C1), ensure that human moderators are available whenever users can livestream using the service	All user-to-user services offering one-to-many livestreaming at medium or high risk of terrorism content, grooming, image-based CSAM, assisting or encouraging suicide, hate, or harassment, stalking, threats and abuse offences.	Illegal Content User-to-User Codes
ICU F3: Providers should ensure that users are unable to do the following in relation to a one-to-many livestream by a child in the UK: a) Comment on the content of the livestream; b) Gift to the user broadcasting the livestream; c) React to the livestream; d) Use the service to screen capture or record the livestream; e) Where technically feasible, use other tools outside of the service to screen capture or record the livestream. Providers should use highly effective age assurance to target this measure.	Services that: a) offer one-to-may livestreaming, and b) it is possible for children to access the service, or a part of it.	Illegal Content Users-to-User Codes

Proposed measures	Who should implement this	Relevant Code(s)
Proactive technology		
ICU C11: Providers should assess whether proactive technology to detect or support the detection of target illegal content is available, is technically feasible to deploy on their service, and meets the proactive technology criteria. If so, they should deploy it.	Providers of services that are likely to be accessed by children that are: • large user-to-user services that are medium or high risk for at least one relevant harm • user-to-user services with more than 700,000 monthly UK users that are high risk, for at least one relevant harm • user-to-user services that are file-storage and file-sharing services which identify a high risk of image-based CSAM, regardless of size • All user-to-user services which identify a high risk of grooming	Illegal Content User-to-User Codes
PCU C9: Providers should assess whether proactive technology to detect or support the detection of target content harmful to children is available, is technically feasible to deploy on their service, and meets the proactive technology criteria. If so, they should deploy it.	Providers of services likely to be accessed by children that are: • large user-to-user services that are medium or high risk for at least one relevant harm • user-to-user services with more than 700,000 monthly UK users that are high risk, for at least one relevant harm	Protection of Children User-to-User Code
ICU C12: Providers should assess existing proactive technology that they are using to detect or support the detection of target illegal content against the proactive technology criteria and, if necessary, take steps to ensure the criteria are met.	Large user-to-user services that are medium or high risk for at least one relevant harm User-to-user services with more than 700,000 monthly UK users that are high risk, for at least one relevant harm User-to-user services that are file-storage and file-sharing services which identify a high risk of image-based CSAM, regardless of size All user-to-user services which identify a high risk of grooming	Illegal Content User-to-User Codes

Proposed measures	Who should implement this	Relevant Code(s)
PCU C10: Providers should assess existing proactive technology that they are using to detect or support the detection of target content harmful to children against the proactive technology criteria and, if necessary, take steps to ensure the criteria are met.	Providers of services likely to be accessed by children that are: • large user-to-user services that are medium or high risk for at least one relevant harm • user-to-user services with more than 700,000 monthly UK users that are high risk, for at least one relevant harm	Protection of Children User-to-User Code
Intimate image abuse (IIA) hash matching		
ICU C14: Providers use perceptual hash matching to detect image-based intimate image abuse content so it can be removed.	Providers of user-to-user services which are high risk of intimate image abuse and: • whose principal purpose is the hosting or dissemination of regulated pornographic content;[1] or • are file-sharing and file-storage services; or • have more than 700,000 monthly active UK users. Providers of large[2] user-to-user services that are medium risk for intimate image abuse.	Illegal Content User-to-User Codes
ICS C8: Providers use perceptual hash matching to detect image-based intimate image abuse content so it can be moderated.	Providers of large general search services.	Illegal Content Search Codes

[1] Regulated pornographic content is pornographic content which excludes content which consists only of text or text accompanied by one of more of the following: identifying content consisting only of text, identifying content which is not pornographic, a GIF which is not pornographic, or an emoji or other symbol.

[2] Large refers to any search or user-to-user service with more than 7,000,000 monthly UK users.

Proposed measures	Who should implement this	Relevant Code(s)
Terrorism hash matching		
ICU C13: Providers use perceptual hash matching to detect terrorism content so that it can be removed.	Providers of the following user-to-user services that enable regulated user-generated content in the form of photographs, videos or visual images (whether or not combined with written material) to be generated, uploaded or shared: • Large services at medium or high risk of terrorism content. • Other services which are at high risk of terrorism content and: o Have more than 700,000 monthly active UK users; or o Are file-storage and file-sharing services	Illegal Content User-to-User Codes
CSAM Hash Matching		
ICU C9: Providers should ensure that hash-matching technology is used to detect and remove child sexual abuse material (CSAM).	Extending the measure to providers of user-to-user services which are high risk of image-based CSAM and where the principal purpose of the service is the hosting or dissemination of regulated pornographic content.	Illegal Content User-to-User Codes
Recommender systems		
ICU E2: Providers should design and operate their recommender systems to ensure that content indicated potentially to be certain kinds of priority illegal content is excluded from the recommender feeds of users.	Providers of user-to-user services that have a content recommender system and are medium or high risk for at least one kind of the following kinds of illegal harms: hate, terrorism, suicide, or foreign interference offences.	Illegal Content User-to-User Codes

Proposed measures	Who should implement this	Relevant Code(s)
User sanctions		
ICU H2: Providers should prepare and apply a sanctions policy in respect of UK users who generate, upload, or share illegal content and/or illegal content proxy, with the objective of preventing future dissemination of illegal content.	All user-to-user services	Illegal Content User-to-User Codes
PCU H2: Providers should prepare and apply a sanctions policy in respect of UK users who generate, upload, or share content harmful to children and/or harmful content proxy, with the objective of preventing future dissemination of harmful content.	All user-to-user services, likely to be accessed by children, which prohibit one or more kinds of PPC, PC and/or NDC	Protection of Children User-to-User Code
ICU C4: Providers should set and record performance targets for their content moderation function covering the time period for taking relevant content moderation action.	Large and/or multi risk user-to user-services, where take down is not technically feasible.	Illegal Content User-to-User Codes
PCU C4: Providers should set and record performance targets for their content moderation function covering the time period for taking relevant content moderation action.	Large and/or multi risk user-to-user services, likely to be accessed by children, where content and/or access level action is not technically feasible.	Protection of Children User-to-User Code
CSEA user banning		
ICU H3: Providers should ban users who share, generate, or upload CSEA, and those who receive CSAM, and take steps to prevent their return to the service for the duration of the ban	All user-to-user services	Illegal Content User-to-User Codes

Proposed measures	Who should implement this	Relevant Code(s)
Highly effective age assurance		
ICU B1: A measure that defines highly effective age assurance for the purposes of the Illegal Content User-to-user Codes and sets out principles that providers should have regard to when implementing an age assurance process.	Services that use highly effective age assurance to determine which UK users of the service are child users for the purpose of targeting measures recommended in the Illegal Content User-to-user Codes at such users.	Illegal Content User-to-User Codes
ICU D15: Providers should allow for appeals of highly effective age assurance decisions and take appropriate action where these are upheld.	Services that use highly effective age assurance to determine which UK users of the service are child users for the purpose of targeting measures recommended in the Illegal Content User-to-user Codes at such users and which are either large or multi-risk.	Illegal Content User-to-User Codes
ICU D16: Providers should allow for appeals of highly effective age assurance decisions and take appropriate action where these are upheld.	Services that use highly effective age assurance to determine which UK users of the service are child users for the purpose of targeting measures recommended in the Illegal Content User-to-user Codes at such users and which are neither large nor multi-risk.	Illegal Content User-to-User Codes
PCU B1: Amendments to codify the definition of highly effective age assurance in the Protection of Children User-to-user Code	Services likely to be accessed by children that use highly effective age assurance to target measures in the Protection of Children User-to-user Code at child users.	Protection of Children User-to-user Code

Proposed measures	Who should implement this	Relevant Code(s)
Increasing effectiveness for U2U settings, functionalities, and user support		
ICU F1 & F2: We are recommending amendments to ICU F1 ('safety defaults for child users') and F2 ('support for child uses'). These proposed changes would mean that services in scope of ICU F1 and F2 should implement these measures using one of the following approaches: • Using highly effective age assurance, as set out in the Part 3 HEAA Guidance; or • Applying ICU F1 and F2 to all users of the service.	All user-to-user services at high risk of grooming User-to-user services that have at least a medium risk of grooming, and are a large service (7 million or more monthly UK users)	Illegal Content User-to-User Codes
Crisis response		
ICU C15 / PCU C11:[3] The provider should prepare and apply an internal crisis response protocol. It should also conduct and record a post-crisis analysis. Providers of large services should implement a dedicated communication channel by which law enforcement can contact them on crisis-related matters during a crisis.	Providers of large user-to-user services that are medium risk of relevant harms Providers of all user-to-user services that are high risk of relevant harms	Illegal Content User-to-User Codes Protection of Children User-to-User Code

[3] The proposed measure PCU C11 will only apply to services that are likely to be accessed by children.

Proposed measures	Who should implement this	Relevant Code(s)
Appeals		
Various ICU D measures (‘Reporting and complaints’): Broadening the scope of appeals measures to ensure that they cover decisions taken on the basis that content was an ‘illegal content proxy’	Various (as set out in the amended measures)	Illegal Content User-to-User Codes
Various PCU D measures (‘Reporting and complaints’): Broadening the scope of appeals measures to ensure that they cover decisions taken on the basis that content was a ‘content that is harmful to children proxy’	Various (as set out in the amended measures)	Protection of Children User-to-User Code

The table below sets out a summary of our proposed amendments to the Illegal Content Judgements Guidance (ICJG).

Proposed amendments	Relevant Document
We are proposing to amend our ICJG to guide providers about how to make illegal content judgements about CSAM in circumstances where they are unable to review content within messages, group chats or forums, but are able to review the associated name, icon or bio / description.	**Illegal Content Judgements Guidance**

3. Introduction and our approach to developing our proposed measures

The Online Safety Act 2023 places new duties of care on online service providers to protect their users from harm and to place safety by design at the heart of their services. Ofcom's Codes of Practice set out ways providers can comply with these duties – many of which are not currently being used by the riskiest platforms.

We published our first Codes for Illegal Harms and Protection of Children in December 2024 and April 2025 respectively. We did this as quickly as possible, so we could start driving improvement by service providers – including by taking enforcement action.

Online harms, technology and evidence evolve quickly. We are keeping pace with developments and responding to evidence we have received by proposing additional measures to build on our first-edition Codes. The new measures we are proposing will deliver meaningful improvements to the Codes by making platforms go further to protect users. Our proposals include:

- new measures on crisis response and recommender systems to reduce the risk of illegal content reaching large numbers of people;
- a package of measures to mitigate the risks to children associated with livestreaming; and
- recommending more use of automated technology to make services safer by design.

This chapter starts by setting out the background and statutory basis of Ofcom's role. It then gives an overview how we chose the proposals we are consulting on, and confirms our approach to impact assessment, evidence, and how we decided who our proposed measures should apply to. Finally, it describes how to navigate the remainder of this document.

Introduction

1.2 This is the introduction to Ofcom's Additional Consultation on our Illegal Content Codes and Protection of Children Codes. A consultation is the formal process we go through to gather views and evidence from stakeholders to help inform our decisions.

1.3 This consultation continues our work to implement the new online safety regulatory regime established by the Online Safety Act 2023 (the Act). It should be read alongside our Registers of Risks (Registers)[4], the existing Illegal Content and Protection of Children Codes of Practice[5] (Codes) and associated guidance.

[4] Ofcom, 2024. Register of Risks. Ofcom, 2025. Children's Register of Risks. [accessed 16 June 2025]

[5] Ofcom, 2024. Illegal Content Codes of Practice. Ofcom, 2025. Draft Children's Codes of Practice [accessed 16 June 2025]. The Secretary of State has laid the draft Protection of Children Codes of Practice before Parliament. Neither House of Parliament resolved not to approve the draft Codes. Therefore, Ofcom will issue the final Codes shortly and they will come into force on 25 July 2025. As at 30 June (the date on which we have

1.4 This chapter sets out the statutory basis of Ofcom's role and the issues we must consider when preparing Codes.

1.5 The chapter then discusses the following aspects of our approach:

- our strategy behind our proposals for new measures in the Codes;
- our approach to the impact assessment of measures in the Codes; and
- who our proposals apply to.

Ofcom's duties and online safety functions

Ofcom's general duties under the Communications Act 2003

1.6 Ofcom is the independent regulator for communications services. We have regulatory responsibilities for the telecommunications, post and broadcasting sectors, as well as for online services.

1.7 As a public authority, Ofcom must act lawfully, rationally and fairly.

1.8 The Communications Act 2003 (the 2003 Act) places duties on us that we must fulfil when exercising our regulatory functions, including our online safety functions. The 2003 Act states that our principal duty in carrying out our functions is:

- to further the interests of citizens in relation to communication matters; and
- to further the interests of consumers in relevant markets, where appropriate by promoting competition.[6]

1.9 In performing that principal duty, we must have regard to principles set out in the 2003 Act, which says that regulatory activities should be transparent, accountable, proportionate, consistent and targeted only at cases where action is needed.[7]

1.10 When performing our duties, we must ensure that UK citizens are properly protected from harm caused by content on regulated services. We achieve this by requiring providers to use suitable systems and processes that help minimise the risk of harm.

1.11 The 2003 Act further requires[8] that we must have regard to the following factors as they appear to us to be relevant in the circumstances.[9] In making our decisions, we have considered factors including, but not limited to:

- the risk of harm to UK citizens presented by regulated services;
- the need for a higher level of protection for children than for adults;

published this consultation) the Codes are still in draft form. However, for the remainder of this document, we refer to these Codes as the Protection of Children User-to-user Code of Practice and the Protection of Children Search Code of Practice.

[6] Section 3(1) of the 2003 Act.

[7] We must also have regard to any other principles appearing to us to represent best regulatory practice.

[8] Section 3(4A) of the 2003 Act.

[9] In relation to matters to which section 3(2)(g) is relevant. The 2003 Act 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. They include: the desirability of promoting competition and encouraging investment and innovation in relevant markets; 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 desirability of preventing crime and disorder; the opinions of consumers and of members of the public generally; and the different interests of persons in the different parts of the UK and of the different ethnic communities within the UK. See Volume – Schedule 4 tests.

- the need for it to be clear to providers of regulated services how they may comply with their duties under the Act;
- the need to exercise our functions to secure that providers may comply with such duties by taking or using measures, systems or processes which are proportionate to the size or capacity of the provider and the level of risk (and potential severity) of harm presented by the service;
- the desirability of promoting the use of technologies which are designed to reduce the risk of harm to citizens; and
- the extent to which providers demonstrate, in a way that is transparent and accountable, that they are complying with their duties.

1.12 In line with our additional duties under the 2003 Act,[10] we have also considered the vulnerability of children and of others whose circumstances put them in need of special protection. We have considered:

- the needs of disabled people, older people, and of those on low incomes;
- the opinions of consumers and of members of the public generally;
- the interests of persons in the different parts of the UK; and
- the interests of the different ethnic communities within the UK.

The Online Safety Act 2023 and the Codes of Practice

1.13 The Online Safety Act 2023 (the Act) is a set of laws designed to protect children and adults online. It puts a range of duties on providers of user-to-user and search services, giving them more responsibility for their UK users' safety. The Act imposes new duties on providers to identify, mitigate and manage the risk of harm from illegal content and activity, as well as content and activity that is harmful to children.

1.14 The Act establishes Ofcom as the regulator responsible for online safety. It places a requirement on us to prepare and issue Codes, which are a package of measures recommended for service providers to comply with their safety duties.

1.15 The safety duties in the Act only apply to services with links to the UK.[11] They also only apply to the design, operation and use of the service in the UK, or (where the duties relate to users of a service) the design, operation and use of the service as it affects UK users.[12] Consistent with this, we can only recommend measures which relate to the design or operation of a service in the UK or as it affects UK users of the service.[13]

1.16 In December 2024[14] and April 2025[15], we published two statements which set out the measures we chose to include in the Illegal Content Codes and Protection of Children Codes of Practice. We recommend service providers implement these measures to be compliant with their obligations in the Act.

1.17 Service providers do not need to follow the Codes and may seek to comply with their safety duties by taking what the Act calls "alternative measures". Where providers take alternative measures, they must keep a record of what they have done and explain how the

[10] Section 3(4) of the Communications Act 2003.

[11] This is defined in section 4 of the Act.

[12] Section 8(3) of the Act.

[13] Schedule 4 to the Act, paragraph 11.

[14] Ofcom, 2024. Protecting People from Illegal Harms Online. [accessed 16 June 2025]

[15] Ofcom, 2025.Protecting Children from Harms Online. [accessed 16 June 2025]

relevant safety duties have been met. They should also have particular regard to the importance of the rights of their users to freedom of expression within the law and protecting the privacy of users.

1.18 The Act says the Codes are like a "safe harbour", meaning that service providers who choose to implement all applicable measures in the Codes will be treated as complying with their relevant duties under the Act.

1.19 Under the Act, we are required to prepare and issue the following sets of Codes for Part 3 services i.e. user-to-user and search services. The Codes include:

- a Code covering terrorism content (relating to the offences set out in Schedule 5);
- a Code covering child sexual exploitation and abuse (CSEA) content (relating to the offences set out in Schedule 6); and
- one or more Codes for the purpose of compliance with other relevant duties (including but not limited to those relating to the offences set out in Schedule 7).

1.20 Ofcom has issued four Codes which collectively meet this obligation – the Illegal Content User-to-User Codes, the Illegal Content Search Codes, the Protection of Children User-to-User Code, and the Protection of Children Search Code. Together these Codes cover all the kinds of illegal harm and harm to children captured under the Act.

1.21 This consultation proposes additional safety measures which could form part of these Codes. The measures in this consultation will have an impact on all Codes except the Protection of Children Search Code.

Schedule 4 and specific additional Illegal Content and Protection of Children Codes considerations

1.22 The Act sets out that Ofcom must consider the appropriateness of the measures we recommend to different kinds and sizes of services and to providers of differing sizes and capacities.[16] We must also have regard to the principles that:

- providers must be able to understand which measures apply to their service;
- the measures must be sufficiently clear, and at a sufficiently detailed level, that providers understand what they entail in practice;
- the measures must be proportionate and technically feasible; and
- the measures must be proportionate to our assessment of the risk of harm presented by services of that kind or size.

3.1 We must also ensure that the measures described in the Codes are compatible with the pursuit of the list of online safety objectives set out in Schedule 4, and that we include measures relating to each of the areas specified in sections 10(4) and 27(4) (concerning illegal content) and sections 12(8) and 29(4) (concerning content harmful to children).[17]

[16] Schedule 4 of the Act.

[17] For user-to-user services, these areas include design of functionalities, algorithms and other features; 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; and content moderation, including taking down content. For search services, these areas include design of functionalities, algorithms and other features relating to the search engine; and content prioritisation. We discuss the areas specified in sections 10(4), 12(8), 27(4) and 29(4) of the Act in more detail in Annex 16 - Legal Framework.

1.23 Under the 2003 Act, we are also required to conduct impact assessments when preparing a Code or amendment to a Code, including an assessment of the impact on small and micro businesses.[18]

1.24 We consider that assessing measures based on our impact assessment criteria is the right approach to ensuring the Codes protect UK users from illegal content and content harmful to children online while also protecting their rights and enabling service providers to operate and innovate in the market.

Human rights

1.25 Our approach to assessing the impact of our measures on human rights is the same as our approach set out in our Statement on Protecting People form Illegal Harms Online (December 2024 Statement) and the Statement on Protecting Children from Harms Online (April 2025 Statement).[19] In this section, we summarise that approach.

Rights relevant to our proposals

1.26 It is unlawful for Ofcom to act in a way that is incompatible with the European Convention on Human Rights (ECHR).[20]

1.27 Of particular relevance to Ofcom's functions under the Act are the right to freedom of expression (Article 10 ECHR) and the right to privacy (Article 8 ECHR). In formulating our proposals in this consultation, we have analysed where we have identified the potential for interference with ECHR rights to make sure any such interference is proportionate.

1.28 The right to freedom of expression includes the freedom to hold opinions and to receive and impart information and ideas without interference by public authority. Article 10(2) of the ECHR states that this right may be restricted in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.

1.29 Article 8(1) of the ECHR states that everyone has the right to respect for his private and family life, his home and his correspondence. Article 8(2) sets out limited qualifications, stating that public authorities must not interfere with the exercise of this right unless necessary in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

1.30 Other ECHR rights which may also be relevant to Ofcom's functions under the Act are the right to freedom of thought, conscience and religion (Article 9 ECHR) and the right to freedom of assembly and association (Article 11 ECHR).

1.31 These are qualified rights. The need for any interference with these rights must be construed strictly and established convincingly. Any interference must be prescribed by or

[18] Section 7 of the 2003 Act, as amended by section 93 of the Act.

[19] Ofcom, 2024. Introduction, our duties, and navigating the Statement, from paragraph 1.15; and Our Approach to developing Codes measures, from paragraph 1.97; Ofcom, 2025. Volume 4 What should services do to mitigate the risks of online harms to children, from paragraph 10.58 [accessed 13 June 2025]

[20] Section 6 of the Human Rights Act 1998.

in accordance with the law, pursue a legitimate aim,[21] and be necessary in a democratic society. In other words, it must be proportionate to the legitimate aim pursued and correspond to a pressing social need.

1.32 We note that service providers also have specific duties under the Act to have particular regard, when deciding on and implementing their safety measures and policies, to the importance of protecting users' right to freedom of expression within the law and from a breach of any statutory provision or rule or law concerning privacy.[22] Service providers will need to do this when deciding on whether to implement, and in implementing, any measures recommended in the Codes. In addition, we have clearly indicated in the Codes where we consider measures act as safeguards for freedom of expression or privacy.

Our approach to assessing when and whether it is proportionate to interfere with these rights

1.33 In considering whether impacts on these rights are proportionate, our starting point is to recognise that the UK Parliament has determined that providers of regulated services must take proportionate measures to protect users from illegal content and content harmful to children – and, where relevant, the commission and facilitation of priority offences. We therefore start from the position that UK users should be protected from the harms set out in the Act and place weight on all the specific evidence of harm set out in our Registers of Risks. A substantial public interest exists in these outcomes.

1.34 We recognise that online safety regulation may also help protect individuals' human rights. For example, when users (including children) feel safe online, they may be more able to exercise their rights to freedom of expression.

1.35 We note that protecting victims' and survivors' human rights is implicit in our duty to carry out our functions so as to secure the adequate protection of citizens from harm presented by content on regulated services.[23] As noted in our December 2024 Statement, we do not consider it necessary to show that a particular harm to a user infringes their human rights in order to show that the user should be protected from that harm.[24]

1.36 Our assessment of human rights in our impact assessment focuses on whether there is reason to think that a measure which would be effective to address the harm amounts to a disproportionate interference with human rights. In order to assess this, we need to consider the impacts of each measure on the rights that are being interfered with.

1.37 We also note that the intent of the Act to ensure that children should be afforded a higher level of protection than adults and should be protected from content that is harmful to them.[25] Our assessment of human rights, including the right to freedom of expression and privacy, has regard to this principle. We have had to consider the impacts of each measure on the rights that are being interfered with, including the rights to freedom of expression and privacy, in order to understand whether a measure – which we think is effective in

[21] As set out in Articles 8(2), 9(2), 10(2) and 11(2). The relevant legitimate aims that Ofcom acts in pursuit of in the context of our functions under the Act include the prevention of crime and disorder, public safety and the protection of health or morals (including those of children), and the protection of the rights and freedoms of others (including children).

[22] Section 22 and 33 of the Act.

[23] Section 3(2)(g) of the 2003 Act.

[24] Ofcom, 2024 Our Approach to developing Codes measures, paragraph 1.99. [accessed 16 June 2025]

[25] Section 1 of the Act.

addressing risks of harm to children – would disproportionately interfere with these rights – in other words, whether the level of interference is no more than needed to secure the objective.[26] Where we have found a measure is likely to interfere with these rights, we have explained why we consider this is proportionate to the Act's legitimate objective of protecting children from harmful content, in which there is a substantial public interest.

1.38 Overall, we have sought to strike a fair balance between securing adequate protections for users from harm (and their human rights in respect of this) and the ECHR rights of users, other interested persons (including for example, persons who host websites or who may be featured in content on regulated services or whose content might be on those services regardless of whether or not they are service users), and service providers, as relevant.

Service providers' rights to decide to remove content that is not illegal or harmful to children

1.39 We note that a service provider has the right to decide to remove content that is not illegal or harmful to children.[27] This is an exercise of its own right to freedom of expression. We cannot compel a provider to carry content it does not wish to carry, nor can we prevent a provider from taking down content that is not illegal or harmful. We acknowledge the risk that, as a result of our recommendations, a provider may choose to take action against content that is legal in order to ensure that it is compliant with its duties relating to content that is illegal or harmful to children. There may be some risk of providers choosing to err on the side of caution, resulting in 'over-moderation' or action on content which would not fall in scope of the definition of illegal content or content harmful to children under the Act. We are required by the Act to provide guidance on the types of content that we consider to be harmful to children and guidance to support service providers in understanding their regulatory obligations when making judgments about whether content is illegal.[28] We consider that this will help providers understand the kinds of content in relation to which they are required to act. Where possible, we have also sought to mitigate this risk. We have drawn out clearly in our Codes where we consider measures act as safeguards for human rights. However, the risk is ultimately one which arises from the scheme of the Act and cannot be mitigated entirely.

Domestic laws relevant to the right to privacy

1.40 Along with the right to privacy conferred by Article 8 ECHR, there are various domestic laws relevant to this right. Service providers will need to ensure they comply with UK 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 2003 (PECR).

1.41 Users' rights to data protection are regulated by the Information Commissioner's Office (ICO). The ICO has a range of data protection compliance guidance which we encourage service providers to consult. In particular, providers should familiarise themselves with the

[26] We note that this is also a requirement under paragraph 10 of Schedule 4 to the Act which require measures to be designed in light of the principles of the importance of protecting the rights of users and interested persons to freedom of expression and privacy.

[27] Ofcom, 2024. Our Approach to developing Codes measures, paragraph 1.104; Ofcom 2025, *Volume 4 What should services do to mitigate the risks of online harms to children, paragraph 10.68 [accessed 17 June 2025]

[28] Ofcom, 2024 Illegal Content Judgements Guidance (ICJG); Ofcom , 2025 Guidance on content harmful to children, [accessed 17 June 2025]

ICO's Children's Code,[29] the ICO's Opinion on Age Assurance,[30] and the ICO Guidance on Online Safety and Data Protection[31] which includes its guidance on content moderation and data protection.[32]

1.42 In our impact assessments, we have explained where we consider that data protection laws may be engaged. We have designed the measures on the basis that those laws will apply.

Equality impact assessment and Welsh language

1.43 We have considered the equality impacts of our proposals, detailing our understanding of any particular impacts on protected groups in the UK.

1.44 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 the need to treat the Welsh language no less favourably than English (in accordance with Welsh language standards).

1.45 We have set out our considerations on these matters in our Equality Impact Assessments at Annex 1.

The government's strategic priorities for online safety

1.46 Ofcom's duties in relation to statements of strategic priorities are set out in section 92 of the Act. We must have regard to a designated statement of strategic priorities for online safety when carrying out our online safety functions, must explain in writing how we propose to do this within 40 days of the statement being designated (or such longer period the Secretary of State may allow), and must publish a review every year of what we have done.

1.47 At the time of writing, the Statement of Strategic Priorities has completed its Parliamentary passage, and we expect DSIT to formally designate is shortly. Ofcom will be responding formally to the SSP in due course.

Ofcom's strategy for additional proposals

Iteration

1.48 We have been clear throughout the process of implementing the Act that our regulatory approach must be dynamic. The nature of UK users' lives online is constantly evolving with the development of new technologies and services. This can result in the emergence of new manifestations of harm.

1.49 Our priority has been to ensure that UK citizens can experience a safer life online as soon as possible. We have moved to implement the Codes as quickly as possible while recognising that we will need to continue to iterate them.

1.50 Accordingly, we took the decision to bring forward an additional consultation and have progressed work on further measures at the same time as rolling out the rest of the regulatory framework.

[29] ICO, Age appropriate design: a code of practice for online services | ICO. [accessed 13 June 2025]

[30] ICO, Age assurance for the Children's code | ICO [accessed 13 June 2025]

[31] ICO, Online safety and data protection | ICO [accessed 13 June 2025]

[32] ICO, Content moderation and data protection | ICO [accessed 13 June 2025]

1.51 We undertook a prioritisation exercise to identify measures to bring forward as part of this consultation process. In this exercise, we considered factors including:

- **Strategic alignment:** We have set out targets for immediate action from industry, by assessing where harm to UK users is greatest and where there are clear steps services can take to protect them. We have selected measures which demonstrate strong alignment with these priorities, and considered how they would enhance the impact of our existing measures. For example, we consider that our proposals on automated content moderation will reinforce existing measures on reporting and content moderation, making services safer by design by filtering out more illegal content before it ever reaches UK users.
- **Evidence of harm:** We have set out measures which we consider will have a significant additional impact on online safety in line with our overarching goal to deliver a safer life online for UK users. For this reason we have prioritised a number of measures relating to the risks of CSEA (which is a priority harm in our Online Safety Roadmap[33]).
- **Stakeholder views:** In developing our proposals, we have had regard to views and evidence stakeholders have presented to us both in responses to previous consultations[34] and in meetings
- **Ability to deliver:** Recognising the imperative to move forward quickly, we have prioritised measures where we were confident that we could produce an evidenced, impact-assessed proposal for the timelines of this consultation.
- **Our obligations and the legislative framework:** In all cases, our ability to act is bound by the Act and other relevant legislation.

1.52 As a result of this exercise, we are now proposing the measures set out in Chapter 2. We are of the view that these measures will make a significant difference to the safety of UK users. Specifically, they will protect users by putting more requirements on providers to protect them from the harm associated with illegal content and content harmful to children.

1.53 The measures support the intent of the Act as set out in our Online Safety Roadmap. For example:

- Our proposals on automated content moderation and recommender systems will make platforms safer by design by meaning less illegal or content harmful to children reaches UK users.
- User banning and applying age assurance more widely will help protect children from sexual exploitation.
- The introduction of hash-matching for intimate image abuse will help protect women and girls online.
- Crisis response measures will ensure services take steps to slow or prevent the spread of illegal material or content harmful to children during periods of high risk.

1.54 This is not the only opportunity to update the Codes or consult on future measures. We will continue to review the effectiveness of the regulatory framework as it comes into force and

[33] Ofcom, Implementing the Online Safety Act: Progress update

[34] November 2023 Consultation: Consultation on Protecting People from Illegal Harms Online. May 2024 Consultation. Consultation on protecting Children from Harms Online. August 2024 Further Consultation Illegal Harms Further Consultation: Torture and Animal Cruelty [accessed 16 June 2025]

we begin to exercise our enforcement powers. Our first enforcement cases are now underway.

1.55 Our upcoming next steps in the implementation of the regime will include a consultation on the Fraudulent Advertising Code, and a further consultation on duties for categorised services. We will also update our Online Safety Roadmap later this year. This will include details of our proposed approach to iterate and update the Codes in the longer term.

Our approach to assessing measures via our impact assessment framework

1.56 Our approach to the impact assessment of the measures in this consultation remains largely as set out in Chapter 2 of the December 2024 Statement, 'Our Approach to Codes'.

1.57 In summary, when assessing the case for including a measure in Codes, we have considered the following factors:

- the prevalence and impact of the harm the measure is combatting;
- the efficacy of the measure in combatting this harm;
- the direct and indirect costs of the measure;
- the impact the measure would have on privacy and freedom of expression; and
- any risks associated with the measure.

1.58 We consider all these factors to assess whether the measure in question is proportionate. As part of this assessment, we consider whether it would be proportionate to apply the measure to all service providers or to a subset of service providers. We also consider both the individual and cumulative impact of the measure.

1.59 While we may have evidence of risk of a harm (or of how it manifests), we do not always have evidence about effective ways to proportionately mitigate this risk (as is needed when making recommendations in the Codes). For example, we do not always have evidence of which measures are effective or what unintended consequences they may have.

1.60 The challenges are bigger when potential measures have significant human rights impacts (as is the case wherever a measure may involve surveillance, user banning, or the use of proxies for illegal content). They are also acute where potential measures involve the use of proactive technology, as the Act sets out extra factors to which we must have regard before we can make recommendations. Unless a particular measure is directly taken from a duty under the Act, we have made recommendations only where we consider we can justify doing so.

1.61 Compared to the December 2024 Statement, the key difference in this consultation is that the Codes are now in place, which means that the baseline for our assessment is different. Accordingly, we are now assessing the anticipated impacts, costs, and benefits of measures over and above those which we expect to result from existing measures.

1.62 The cumulative impact assessment (which is included in Chapter 22) captures the additional costs and benefits expected to result because of all the measures included in this consultation.

Our use of evidence

1.63 As an evidence-based regulator, we continue to use information from a diverse set of sources. We have sought the best available information, irrespective of whether that information has come from industry, civil society, research, or other sources, including through previous consultations.

1.64 Where necessary and feasible to develop a proposal for consultation, we are using our powers in the Act to make a request for information from online services to provide information on a specific topic.

1.65 In setting out our proposals and our decisions we have used a range of evidence, including:

- responses to our July 2022 and January 2023 Calls for Evidence;[35] [36]
- responses to our November 2023 Consultation, our May 2024 Consultation, and our August 2024 Illegal Harms Further Consultation on Torture and Animal Cruelty (August 2024 Further Consultation);
- information gathered through bilateral sessions and other sessions with stakeholders, including engagement with industry groups, civil society groups, groups representing the interests of individuals with lived experience of online illegal harms, and harms to children and our deliberative engagement with children;
- our own and others' research, including research on evidence around the kinds of and risk of harm online, as well as important areas across the Codes and wider statement documents;
- we have tried, where possible and appropriate, to incorporate learnings from our Video Sharing Platform work and align with international counterparts.

1.66 Online service providers within the scope of the Act (and the technologies they use) are evolving rapidly, and new harms may emerge as a result. There is a need for prompt action to protect people online. Therefore, some of our proposals are based on an assessment of more limited or indirect evidence of impact and have a reliance on logic-based rationales. We welcome comments on this approach, as well as additional evidence in relation to any of our proposals in this consultation.

Who these measures apply to

1.67 We have continued the approach taken in the December 2024 Statement and the April 2025 Statement, in that each proposed measure includes details about the services it should apply to. This aligns closely with the requirement for our proposed measures to be proportionate and acknowledges the diversity of services that fall in scope of the measures.

1.68 In some cases, including where required by the 2023 Act, we propose different measures for user-to-user and search services. We also distinguish between general and vertical search services. Beyond that, we have broken down services in scope of the Act into a number of broad types:

- We define services which are medium or high risk for a particular harm as 'single risk' services;
- We define services that are medium or high risk for two or more kinds of harm as 'multi risk' services;
- We define large services as being those with an average user base greater than seven million monthly UK users. We define 'UK users' in line with the definition used in the

[35] Ofcom, 2022: Call for Evidence: First phase of online safety regulation.

[36] Ofcom, 2003: Call for Evidence: Second phase of online safety regulations: Protection of Children

Act for the purposes of determining whether a service is 'large'.[37] The average monthly number of UK users should be calculated over six months.[38] Our approach of taking user base as a proxy for the size of a service was similar to that adopted by the European Union in the Digital Services Act.[39]

1.69 In taking decisions about the scope of our proposals in this consultation, we have followed the principles set out in the December 2024 Statement, 'Our Approach to Codes'.[40]

1.70 Our main focus is the extent to which measures can reduce risks to people in the UK; based on the available evidence of harm and our current understanding of the most effective ways to mitigate this. The Act requires us to ensure measures are proportionate, and we recognise that the size, capacity, and risks of services differ widely. We therefore do not take a one-size-fits-all approach. Instead, we have set out what types of service we think should use specific safety measures to comply with their duties, with the most extensive expectations placed on the riskiest services.

1.71 The size of a service and its user base is one indicator of risk. However, there are some services which are inherently risky even when their reach is small (for example, because they feature risky functionalities or are focused on content which is associated with higher levels of illegality or harm) [41]

1.72 Therefore, some of our proposals in this consultation apply to all providers of high-risk services, whether they are small or large. At the same time, we recognise that the largest services have significantly more extensive reach, and it is therefore reasonable to expect more of the largest providers. We have sought to minimise the regulatory burden on small low-risk services.

How to use and navigate this document

1.73 It is important that this consultation is accessible and clear to navigate for a range of audiences. For this reason, we have:

- used a structure for the main chapters of the document which will be familiar to readers of our previous online safety consultations;
- cross-referenced material from our existing regulatory documents where appropriate with suitable links; and
- ensured each chapter includes a summary setting out what the chapter is about, our proposals, and what information we are seeking from stakeholders.

1.74 We will also engage with relevant stakeholders following publication of this consultation.

1.75 This consultation is broken down into one main volume setting out our proposals, followed by a number of annexes.

[37] A user is a 'UK user' of a service if a) where the user is an individual, they are in the UK; or b) where the user is an entity, it is incorporated of formed under the law of any part of the UK. Where the Act refers to users, it does not matter whether a person is registered to use a service. See section 227(1)-(2) of the Act.

[38] See section 5 of each our Codes of Practice.

[39] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[40] Ofcom, Our Approach to developing Codes measures [accessed 16 June]

[41] For such services, we therefore consider it consistent with our risk-based approach to include all services with a particular functionality as in scope for certain measures. This is the case for livestreaming services

- The first part of the main volume is this chapter. It summarises our proposals, sets out how our consultation links to the overall strategy to implement the online safety regime, and sets out our approach to developing our proposals.
- The remainder of the volume sets out in detail our rationale for proposing the measures included in our consultation. It includes our impact assessments, rights assessments, and detailed questions for stakeholders.

1.76 We have grouped our proposals in this consultation thematically to make it easy to see how measures relate to each other, where appropriate. The chapters are:

- Chapter 4-7: Livestreaming measures
- Chapter 8-13: Automated Content Moderation measures
- Chapter 14: Recommender Systems measures
- Chapter 15-17: User Sanctions measures
- Chapter 18-19: Approach to HEAA and User-to-user safety settings
- Chapter 20: Crisis response measures
- Chapter 21: Broadening Appeals measures
- Chapter 22: Combined Impact Assessment
- Chapter 23: Statutory Tests

1.77 We also have the following annexes:

- Annex 1: Equality Impact Assessments
- Annex 2-5: Responding to this consultation
- Annex 6: Draft Guidance to Proactive Technology Measures
- Annex 7-11: Addenda to regulatory products
- Annex 12: Overview of proposed changes to Codes of Practice
- Annex 13: Proactive Technology: Further evidence on relevant harms
- Annex 14: Perceptual Hash Matching for Intimate Image Abuse
- Annex 15: Assumptions on costs and further analysis on costs and benefits
- Annex 16: Legal Framework
- Annex 17: Glossary

Definition Box: Illegal content proxy

Throughout this consultation we use the term illegal content proxy, we define this as:

For User-to-user services: content that has been assessed and identified as being in breach of the service’s terms of service, where the provider is satisfied that the terms in question prohibit the types of content that include illegal content (including but not limited to priority illegal content).

For search services: search content that has been identified in the provider’s publicly available statement for the service as being subject to appropriate moderation action, where the provider is satisfied that illegal content is included within that kind of content (including but not limited to priority illegal content).

Next steps

1.78 This consultation will remain open until 20 October 2025. During the consultation period, Ofcom will engage with stakeholders on our proposals.

1.79 When our consultation closes, we will consider all the responses and evidence received and will use this to make our decisions on whether and how to proceed with our proposals. In due course, we will publish a statement that sets out our decisions, the rationale behind them, and what changes we intend to make to the Codes as a result.

1.80 The statement and the revised Codes will subsequently need to be laid before and approved by the UK Parliament before coming into force.

4. Livestreaming – Introduction

Warning: this chapter contains content that may be upsetting or distressing

Summary

Livestreaming has both benefits and risks. Users often use one-to-many livestreams for the purposes of entertainment or education – such as streaming gameplay, showcasing talents, citizen-journalism or sharing real-world experiences – and to interact or connect with others.

However, due to its real-time, ephemeral nature, livestreaming poses significant risks – particularly to children. These risks are amplified by supporting functionalities such as the ability to post comments and send gifts. For example, research by the NSPCC shows that 6% of children who livestream have been asked to remove their clothing.[42] Similarly, there are numerous examples of violent attackers using livestreaming to broadcast and share footage of atrocities.

We have existing measures in our Codes of Practice which should help make livestreams safer. We already say that user reporting features should be easy to find, access and use; that providers effectively prioritise content awaiting moderation; and that content moderation teams should be appropriately resourced.

However, there are two sets of circumstances about which we are particularly concerned, and which warrant further action. The first is users creating livestreams specifically to broadcast illegal content, in particular content where there is an imminent risk of physical harm. The second is children being subject to a range of harms while livestreaming, including being groomed, coerced into performing sexual acts, encouraged by other users into acts of self-harm and suicide, and being exposed to comments of a hateful and abusive nature.

We are therefore proposing to add some additional measures to our Codes of Practice to deal with the risks of livestreaming, making it safer by design. We propose that, on all services offering livestreaming as a functionality:

- Users should be able to easily report livestreams that depict imminent physical harm.
- Human moderators should be available to review content and take action in real time whenever livestreaming functions are available.

In addition, where children in the UK are able to livestream, users should not be able to interact with children's livestreams, including by posting comments or reactions or sending gifts, and should be prevented from recording the content of children's livestreams. Services should use highly effective age assurance to apply this measure.

We are not proposing new measures to restrict children's ability to view livestreamed content as we expect this to be addressed to some degree through existing measures in our Codes. However, we anticipate that our proposals will also reduce children's exposure to illegal or content harmful to children by helping to reduce its availability.

We believe these proposals create a baseline level of protection to deal with the risks associated with livestreaming. If we receive compelling evidence during the consultation, we are prepared to go further, which could include recommending that children are prevented from livestreaming entirely.

[42] NSPCC, 2018. Livestreaming and video-chatting, p.1. [accessed 3 June 2025].

Consultation Questions

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

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

What is livestreaming?

4.1 Livestreaming is a user-to-user service functionality that allows users to simultaneously create and broadcast online streaming media in, or very close to, real time. Our previously published advice to the Secretary of State regarding the threshold conditions for categorised services[43] splits livestreaming into one-to-many, one-to-one and many-to-many broadcasts. This chapter – and our proposed measures – focus on one-to-many livestreams because we have identified stronger evidence of harm occurring within these settings.

4.2 For the purposes of our proposed measures, we define a one-to-many livestream as a user-to-user service functionality that allows a single user to simultaneously create and broadcast online streaming content (such as live video content) to multiple other users in, or very close to, real time. Examples include a user livestreaming to their followers on a social media service, or a gamer streaming their gameplay on a dedicated livestreaming service.

4.3 One-to-many livestreaming does not include instances when multiple users are streaming content with each other simultaneously, for example as part of a group chat or a multi-player game. One-to-many livestreams may be visible to, and able to be accessed by, a large number of users. This may include a user's followers, friends or connections, or it may include all users or viewers of a service.

4.4 People livestream in large numbers. Ofcom research suggests 16% of 3- to 17-year-olds and 21% of 13- to 17-year-olds (around 1 in 5) broadcast their own livestreams.[44] Among those aged 18 or over, 21% of men and 13% of women report having broadcast livestream content.[45] It should be noted that no specific definition of livestreaming was provided to respondents to this survey, and so the figures are likely to encapsulate all types of livestreaming, potentially including video calls.

4.5 Livestreaming is used for many social, entertainment (for example gaming), and commerce purposes.[46] Other uses may include citizen journalism, social activism, and real-time crisis reporting. Livestreaming also allows for the development of and engagement with communities of common interest. Identified benefits to livestream broadcasters include engagement with and receiving support from others, immediacy of interaction and

[43] Ofcom, 2024. Categorisation: Advice submitted to the Secretary of State. [accessed 10 June 2025].

[44] Ofcom, 2025. Children and Parents: Media Use and Attitudes Report. [accessed 4 June 2025].

[45] Ofcom, 2025. Adults' Media Use and Attitudes Report. [accessed 4 June 2025].

[46] Huang, Z., Mou, J., Benyoucef, M. and Kim, J., 2023. Live Streaming: Its Relevant Concepts and Literature Review. [accessed 4 June 2025].

feedback,[47] development of creative, entrepreneurial, public speaking and communication skills and confidence, and financial opportunities.[48]

4.6 However, livestreaming also poses multiple risks to users, as detailed in this chapter. Many of these are linked to the instantaneous and transitory nature of the content and interactions that take place, and challenges in moderating livestreamed content. This makes it riskier than, for example, producing and uploading a video of similar content. One-to-many livestreaming also tends to be accompanied by supporting functionalities. These may include the ability for users watching the livestreams to post comments, send gifts, and react in different ways such as through likes, dislikes, and emojis. These functionalities can pose risks and enable various forms of harm, particularly to children.

4.7 In relation to children, given the severity of some of the risks identified, it is possible that access to livestreaming needs to be made available to them with considerable safeguards, as proposed in our measures. This would have implications for children's ability to enjoy in full the benefits of being able to livestream their own content. Our proposals are not intended to unduly preclude children from viewing livestreams, subject to services' requirements under existing Illegal Content and Protection of Children User-to-user (U2U) Codes of Practice (Codes). Through this consultation and in parallel with it, we are working to obtain further insights on the benefits and risks of livestreaming from children and those that care for them, experts, and those with lived experience of online harms. This will help inform our measures to ensure children can, if possible, continue to benefit from this functionality while being protected from harm.

4.8 In the course of preparing this consultation, we have engaged with those with lived experience of online harms. They expressed a range of views on measures to mitigate harms to children when they broadcast a one-to-many livestream.[49] While these views are not fully representative, they convey a clear desire to ensure that children can continue to enjoy the benefits of technology, but that the onus should be on services to ensure that children are safe and protected while online. Views expressed included:

a) One-to-many livestreaming should not be permitted for children, particularly those aged under 16.

b) Services should monitor comments to detect illegal or inappropriate behaviour. Alternatively, rather than allowing freetext, comments could be limited to a set of specific sentences and questions, such as initial greetings and asking how the livestreamer is, or what they are or have been doing.

c) Emojis could be limited to preset ones which do not have harmful connotations or coercive purposes.

d) Gifting should not be permitted when children are livestreaming.

e) Services should publish transparency data on how they address livestreaming harms and the effectiveness of their efforts.

[47] Huang, Z., Mou, J., Benyoucef, M. and Kim, J., 2023.

[48] Anecdotal evidence of the purposes and benefits of livestreaming for children has been gathered from the following sources: EE, 2025. Your Kid Wants to be a Streamer? How to Help Them 'Go Live' Safely. [accessed 4 June 2025]; Integrative Psych, 2025. Navigating Twitch for Kids: Balancing Benefits and Concerns. [accessed 4 June 2025]; Internet Matters, 2025. What is live streaming & vlogging?. [accessed 4 June 2025]; NSPCC, 2025. Livestreaming and online video apps. [accessed 4 June 2025]; Streamer Facts, 2024. How much do Twitch streamers make in 2025? Small vs Big vs Top-tier [accessed 4 June 2025].

[49] This engagement took place on 30 April and 22 May 2025.

4.9 Those with lived experience of online harms noted some purposes for which children broadcast one-to-many livestreams, including the desire to fit in with friends, share their interests with others, and to livestream aspects of their daily life.

Harmful scenarios we are particularly concerned about

4.10 There are two distinct scenarios that we are particularly concerned about in relation to livestreaming:

- Users creating livestreams to broadcast illegal content, in particular content where there is an imminent risk of physical harm. This includes terrorist, violent and hateful attacks, content which encourages or assists suicide[50] or other livestreams of actual, or risk of, imminent physical harm or violence. This may also include viewers encouraging or inciting the user towards suicide, terrorism or violence using comment functionality.
- Children being subject to a range of harms while livestreaming, enabled by viewers' ability to post comments and reactions, send gifts, and record the content of their livestreams. These harms include being groomed, coerced into performing sexual acts, encouraged by other users to acts of self-harm and suicide, and being exposed to comments of a hateful and abusive nature. We set out various proposed measures to mitigate these risks.

4.11 Children can also be exposed to illegal content and content that is harmful to them as viewers of livestreams. While we are not proposing specific new measures to restrict children's ability to view such content, livestreaming services are already subject to requirements under existing Illegal Harms and Protection of Children Codes, for example regarding content moderation, which should result in a reduction in illegal content and content harmful to children being livestreamed.

4.12 The proposals set out in this chapter do not directly address adult-controlled, commercial livestreaming of child sexual exploitation and abuse. This form of offending typically involves users making payments to adults based outside of the UK, who then facilitate or commit livestreamed sexual abuse of children based abroad in return. The UK has been estimated to be the world's third largest source of payment for livestreaming of such offending from the Philippines.[51] Livestreams of child sexual abuse are priority illegal content; as such, user-to-user service providers have duties to prevent users from encountering it and to take it down once they become aware of it. Our Codes contain measures that we recommend service providers adopt to comply with these duties, including measures related to complaints and content moderation. The scope of these duties (and our measures) is limited to the design, operation and use of the service in the UK and as it affects UK users.[52] Therefore, while this abuse is clearly abhorrent, to the extent that children located overseas are harmed in the creation of these kinds of livestreams, this is not something that the Act can address. Rather, this is primarily a matter for law enforcement (the vast majority of the accounts distributing this abuse are operated by criminal gangs) and for providers of the related financial payment mechanisms

[51] See paragraph E1.5 of: Independent Inquiry Child Sexual Abuse, 2020. The Internet: Investigation Report. [accessed 10 June 2025]. See also paragraph 2.39 of: Ofcom, 2024. Illegal Harms Register of Risks. [accessed 10 June 2025] for more detail on this type of offending.

[52] Section 8(3) of the Online Safety Act 2023.

to address. The focus of the proposals in this chapter is on the harms that may affect children in the UK while they are livestreaming.

4.13 Stakeholder responses to our November 2023 Consultation on Protecting People from Illegal Harms Online (November 2023 Consultation) and our May 2024 Consultation on Protecting Children from Harms Online (May 2024 Consultation) noted the risks posed by livestreaming, and called upon us to implement measures to help address them.

- The Community Security Trust and Antisemitism Policy Trust called on us to do more about livestreamed terror attacks and hate content.[53]
- Barnardo's and the WeProtect Global Alliance noted how livestreaming services often have few or no measures in place to detect livestreamed child sexual abuse or exploitation.[54]
- Multiple organisations highlighted the greater risk of grooming and exploitation that children face from livestreaming.[55] The 5Rights Foundation proposed livestreaming functionality should be 'off' by default for child users, and suggested the use of highly effective age assurance or consideration of policies to prevent livestreams featuring children in bedrooms, classrooms and bathrooms. 5Rights also noted the risk from comment functionality on livestreams, as well as the risk to children from in-game and in-app gifting more generally and proposed that the latter should also be set to 'off' by default.[56]
- The National Society for the Prevention of Cruelty to Children (NSPCC) and the Independent Inquiry into Child Sexual Abuse (IICSA) Changemakers noted evidence from the Internet Watch Foundation (IWF) that the child sexual abuse material (CSAM) it removes online is often derived from livestreams. They highlighted the importance of preventing child sexual exploitation and abuse (CSEA) via livestreaming to disrupt this criminality. Given its high-risk nature, the NSPCC stated that it would expect a future Code of Practice to set out measures requiring companies to tackle CSEA on livestreaming services.[57]
- The IWF noted livestreaming as a high-risk functionality, and suggested that if such risks cannot be sufficiently mitigated, services should prevent all children or children in certain age groups from accessing the functionality.[58]
- The Molly Rose Foundation noted that livestreaming poses particular risk in respect of suicide and self-harm, especially to children.[59] The Online Safety Act Network (OSAN)

[53] Antisemitism Policy Trust response to November 2023 Consultation, p.6-7.

[54] Barnardo's response to November 2023 Consultation, p.11; WeProtect response to November 2023 Consultation, p.3.

[55] 5Rights Foundation response to November 2023 Consultation, p.28.; Cybersafe Scotland response to November 2023 Consultation, p.1; International Justice Mission response to November 2023 Consultation, entire document; NSPCC response to November 2023 Consultation, pp.25-26.; Online Safety Tech Industry Association response to November 2023 Consultation, p.2.; WeProtect response to November 2023 Consultation, pp.2-3.

[56] 5Rights Foundation response to November 2023 Consultation, p.28; 5Rights Foundation Response to May 2024 Consultation, p.6-7

[57] NSPCC response to November 2023 Consultation, p.25-26.

[58] Internet Watch Foundation response to May 2024 Consultation, p.4.

[59] Molly Rose Foundation response to November 2023 Consultation, p.25.

recommended that we create more friction around livestreaming services and child users.[60]

Risks associated with livestreaming

Factors that make livestreaming particularly risky

4.14 From our analysis, we assess that the following factors make livestreaming particularly risky.

Challenges in moderating content

4.15 The real-time nature of livestreaming makes content moderation difficult, even with automated tools. Where harm is identified, moderation technologies and processes may struggle to address the content sufficiently quickly,[61] leaving livestreamers and viewers exposed to harm. There are questions about the extent to which existing moderation processes enable prompt, appropriate action to be taken, including safeguarding of livestreamers and reporting of identified criminality such as grooming.[62]

Real-time interaction can pose risks to livestream broadcasters

4.16 The opportunity for real-time interaction between livestreamers and their viewers amplifies the potential for harm to livestreamers as they are acting 'in the moment'. This can make them more vulnerable to taking decisions or actions that they may not take if they had more time to reflect. This can include doing so for financial reward, validation, or recognition. In some cases, the 'in the moment' nature of the livestream may also lead or motivate users to livestream physical harm or violence against others or towards themselves.

Real-time interaction can also motivate bad actors viewing livestreams

4.17 The 'in the moment' nature of livestreaming can motivate viewers of livestreams to incite and coerce certain behaviours from livestream broadcasters, or to 'pile-on'[63] or attack them in illegal and harmful ways, as they are then able to see the direct consequences of their interactions in real time. This makes it a more appealing and compelling medium for bad actors than pre-recorded content. Such behaviour can range from attempts by single users, to more coordinated and pre-meditated action.

Exposure of viewers of livestreams to harm

4.18 Viewers can also be harmed by seeing livestreaming of illegal content or content harmful to children, and by the actions and interactions of viewers and livestreamers.

[60] Online Safety Act Network response to May 2024 Consultation, p.52.

[61] Cambridge Consultants Report, 2019. Use of AI In Online Content Moderation. [accessed 10 June 2025] states "Whilst live streaming is a powerful tool for connecting users, it raises significant concerns to content moderation systems due to the demanding requirement of analysing complex content with multiple features in real time."

[62] See, for example: CDT Research (Gorwa, R. and Thakur, D.), 2024. Real Time Threats: Analysis of Trust and Safety Practices for Child Sexual Exploitation and Abuse (CSEA) Prevention on Livestreaming Platforms. [accessed 4 June 2025].

[63] 'Pile-on' refers to when a user is criticised or targeted by a large number of other users, often as part of bullying campaigns.

Misperception of the broadcast as temporary

4.19 Some users livestreaming, particularly children, may consider the broadcast as no more than an ‘in the moment’ event. They may not realise that their content could be recorded and shared without consent, which may lead them to act with less restraint. Recording of a livestreamer’s content – particularly where said content is illegal, harmful, or likely to cause embarrassment – could leave them vulnerable to blackmail and extortion and can help to revictimise those that have been harmed or abused when livestreaming, with negative impacts on their wellbeing.

Harms associated with livestreaming

4.20 Livestreaming is listed in Ofcom’s Illegal Harms Risk Profiles and Children’s Risk Profiles respectively as a risk factor for:

- CSEA, incorporating grooming and CSAM; terrorism, encouraging or assisting suicide; hate, harassment/stalking/threats/abuse offences; and animal cruelty; and
- suicide and self-harm content; and violent content.

4.21 Here, and in the subsequent section on risks from supporting functionalities, the evidence we have identified is limited in relation to some types of harm. We welcome further evidence of harms on one-to-many livestreams in response to this consultation, to supplement our own ongoing research.

Child sexual exploitation and abuse (incorporating grooming and child sexual abuse material).

4.22 Livestreaming plays a significant role in grooming and the creation of CSAM, and a significant proportion of children who undertake livestream broadcasts can be subject to grooming attempts. A 2018 study by the NSPCC found that of 40,000 children surveyed, 24% (around 1 in 4) had undertaken a livestream, and of those, 6% had received requests to change or remove their clothing.[64]

4.23 Children’s livestreams are popular with CSEA offenders. An anonymous dark web survey found that 45% (nearly half) of respondents had viewed some form of livestreamed CSAM.[65]

4.24 Various supporting functionalities that typically accompany livestreams are used to groom and coerce children into sexual activity, and to record and share the resulting content, which we outline in more detail in paragraphs 4.36 to 4.75.

[64] Survey of nearly 40,000 children aged 7-16 years old: NSPCC, 2018. Livestreaming and video-chatting, p.1. [accessed 3 June 2025].

[65] Protect Children (Insoll,T., Ovaska, A. and Vaaranen-Valkonen, N.), 2021. CSAM Users in the Dark Web: Protecting Children Through Prevention. [accessed 4 June 2025].

Terrorism

4.25 There are many examples of livestreaming being used to promote terrorism. [66] Livestreaming has been used to broadcast terrorist attacks,[67] [68] with recordings of these attacks often used to incite and encourage terrorism – particularly by extreme far-right terrorists, who often seek to emulate the tactics of previous attackers.[69] While such attacks are relatively rare, and the original content may only be live on a service for a short time, there is evidence that recorded footage is re-shared or forwarded, further radicalising others and retraumatising targeted communities, exacerbating the harm caused.[70]

4.26 The risks associated with such content are linked to the ability of the recorded content to go viral and motivate others to carry out terrorist attacks in a similar manner.[71]

Encouraging or assisting suicide

4.27 Livestreaming functionality is of particular concern in relation to illegal suicide content (meaning content encouraging or assisting suicide).[72] Livestreaming suicide is not in itself illegal. However, livestreamers who are expressing suicidal thoughts, feelings or intent may be in an extremely vulnerable situation, and other users' reactions to or engagement with them, for example via comments, could further encourage an attempt at suicide, and that encouragement may be illegal. The opportunity for real-time interaction between livestreamers and their viewers amplifies the potential for harm to livestreamers as they are acting 'in the moment'. This can make them more vulnerable to taking decisions or actions that they may not take if they had more time to reflect.

4.28 There have been several cases in which one-to-many livestreaming has been used, including by children, to broadcast them self-harming or ending their life in real time,[73] [74] and studies of suicide livestreams have concluded that some comments made by users encouraged the suicide of users who were livestreaming.[75]

[66] Ofcom, 2024. Illegal Harms Register of Risks. [accessed 10 June 2025].

[67] "A gunman went live on a social media service before he shot and killed 51 people at local mosques. In the same year, a gunman in Germany also livestreamed his attack on a social media service." Brooks, A and Matromarino, J.P., 2022. Extremists exploit gaming networks and social media to recruit and radicalize, Wbur, 19 May. [accessed 4 June 2025]

[68] In 2016, a man in France used a social media live feature to broadcast his justification for killing two police officers while holding a child hostage and pledging his allegiance to the Islamic State. In 2019 a gunman reportedly livestreamed himself through his channel on a social media service, attacking a synagogue and a kebab shop in Halle, Germany. In 2020, an attacker livestreamed himself carrying out an attack in a mall in Glendale, Arizona. Ofcom, 2022. The Buffalo attack: Implications for online safety. [accessed 4 June 2025].

[69] Ofcom, 2024. Illegal Harms Register of Risks. [accessed 4 June 2025].

[70] Ofcom, 2024. Illegal Harms Register of Risks. [accessed 4 June 2025].

[71] Ofcom, 2024. Illegal Harms Register of Risks. [accessed 4 June 2025].

[72] Ofcom, 2024. Illegal Harms Register of Risks. [accessed 4 June 2025].

[73] Getz, L., 2017. Livestreamed Suicide on Social Media — The Trauma of Viewership, *Social Work Today*, 17 (3). [accessed 4 June 2025].

[74] For evidence of this in relation to children, see paragraph 3.81 of Ofcom, 2025. Children's Register of Risks. [accessed 12 June 2025].

[75] Phillips, J G. and Mann, L., 2019. Suicide baiting in the internet era. *Computers in Human Behaviour*, 92. [accessed 4 June 2025].

Hate, harassment, stalking, threats and abuse offences

4.29 There is a material risk of livestreams being used to broadcast hateful content with large audiences of users. For example, analysis by OSAN identified the use of livestreaming to incite hate and violence across the UK in July 2024.[76]

Animal cruelty

4.30 Livestreaming can also be a risk factor for animal cruelty, which could potentially include animal fights and torture.[77] For example, the Royal Society for the Prevention of Cruelty to Animals (RSPCA) stated that it had identified livestream broadcasts as a risk factor for dog fights and the feeding of live animals to snakes.[78] The Scottish Society for Prevention of Cruelty to Animals (Scottish SPCA) also stated that livestreaming is a risk factor for animal cruelty, and is sometimes distributed on a pay-to-access basis within the UK.[79]

Harms to children

4.31 In addition to being exposed to illegal harms, there is evidence of children being exposed to content that is harmful to them via livestreams.[80] For example, livestreaming can increase the risk of children being exposed to suicide and self-harm content such as people actively preparing to take their own life or to self-harm.[81] Viewing such content in real time over livestreams can have a more visceral effect on viewers compared to viewing similar pre-recorded content, particularly for children.

4.32 Ofcom research shows that children have been exposed to violent content on livestreams, with the most seen content being livestreams of local school and street fights. The amount of violent content that was livestreamed is not known, but children described encountering violent content online as unavoidable.[82]

4.33 There is evidence associating livestreaming with a risk of exposure to multiple types of content harmful to children. This includes suicide and self-harm, pornographic content, abuse and hate, and content promoting harmful substances. Some of this content and behaviour may be similar to, or overlap with, illegal content – for example, content that encourages suicide, or content that incites hatred. [83]

4.34 The NSPCC noted that children and young people it consulted had highlighted Twitch was a particularly risky service in relation to livestreaming, as when streamers do ‘upsetting or dangerous things’, it is disturbing for any child who sees it. However, it is particularly

[76] Online Safety Act Network, 2024. Disinformation and disorder: the limits of the Online Safety Act. [accessed 12 June 2025].
[77] Ofcom, 2024. Illegal Harms Register of Risks. [accessed 4 June 2025]
[78] RSPCA response to November 2023 Consultation, p.2.
[79] Scottish SPCA response to November 2023 Consultation, pp.2, 4.
[80] The Act requires services to assess and mitigate the risk of harm to children from content that is harmful to them. A definition of harm is provided from paragraph 1.7 of: Ofcom, 2025. Children's Register of Risks. [accessed 12 June 2025].
[81] This could include them discussing their plans in detail or preparing accessories, or tools for suicide or self-harm. See also p.305, Table 8.3.2 and p.311, Table 8.4.2 for descriptions and examples of content which encourages, promotes or provides instructions for suicide and for acts of deliberate self-injury respectively. Ofcom, 2024. Protecting children from harms online - Volume 3: The causes and impacts of online harms to children. [accessed 12 June 2025].
[82] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children. [accessed 4 June 2025].
[83] Ofcom, 2025. Children’s Register of Risks. [accessed 4 June 2025].

dangerous because streamers can have positions of influence and may encourage dangerous behaviour in others.[84]

4.35 We assess that children are likely also to experience a range of harms from comments made on livestreams, including bullying, hate and abuse, or encouraging or providing instruction for suicide or acts of self-harm. We outline these in further detail in paragraphs 4.45 to 4.51.

Risks from supporting functionalities on livestreams

4.36 Our analysis has identified several supporting functionalities which, when used alongside livestreaming, can amplify the risk of harm. These are the ability to comment, send gifts, post reactions, and to record or screen capture the livestream.

4.37 Some of the available evidence does not differentiate the specific livestreaming environments in which the harms take place to enable us to fully isolate evidence of harm on one-to-many livestreams. We have noted where the evidence indicates this is the case, but it is likely that these harms manifest on all types of livestreams to some degree.

Commenting

4.38 Livestreams typically include the ability to post comments as the broadcast is taking place. These comments are usually visible and able to be reacted to immediately by users, particularly when they are registered and logged into a service.[85] We have identified comments on livestreams being used to enable the following harms.

Child sexual exploitation and abuse (including grooming and child sexual abuse material)

4.39 Perpetrators use comments on livestreams to build rapport with victims and survivors in the early stages of grooming. Transcripts from the IWF show that perpetrators leave sexualised comments when children livestream as part of the grooming process, for example to flatter, deceive and manipulate them into producing self-generated CSAM.[86] After obtaining CSAM from children, some perpetrators use comment functionality to extort and blackmail them to produce more, and sometimes more severe, CSAM.[87]

4.40 A 2022 study identified many users of Twitch (a one-to-many livestreaming service) who used comment functionality to seek to manipulate large numbers of children into certain poses and to produce sexual imagery.[88] The Times identified the presence of over 100 cases of sexually explicit comments on children's one-to-many live videos on YouTube in

[84] NSPCC response to May 2024 Consultation, p.25.

[85] The Children's Register of Risks glossary defines commenting on content as "a user-to-user service functionality that allows users to reply to content, or post content in response to another piece of content posted on open channels of communication, visually accessible directly from the original content without navigating away from that content".

[86] King, J. 2022. The shocking transcripts that reveal how groomers sexually abuse children in their own rooms, Metro, 3 September 2022. [accessed 12 June 2025].

[87] Internet Watch Foundation, 2025. Self-generated child sexual abuse. [accessed 28 May 2025].

[88] Bloomberg UK, 2022. Child Predators Use Twitch to Systematically Track Kids Livestreaming. [accessed 9 June 2025].

2018, where perpetrators sought to convince children to take off their clothes or adopt sexualised poses.[89]

4.41 While not specific to livestreaming, perpetrators are also known to use comments on child users' content to share hyperlinks or make attempts to exchange contact details with the aim of getting the child to connect with them on another service.[90]

Encouraging or assisting suicide

4.42 A study of the comment threads of 26 livestreams where an individual had expressed intent to take their own life found that comments attempted to discourage the suicide threat in nearly 9 out of 10 cases (88%).[91] However, in just under half of the cases (11 of the 26), some of the comments encouraged the suicide attempt or insulted the victim. Individuals expressing an intent to take their life on livestreams are likely to be extremely vulnerable to acting on such incitement and encouragement.

4.43 Incitement or encouragement of suicide and self-harm is also becoming increasingly intertwined with CSEA, with the advent of sadistic interest groups or 'Com networks'.[92] These groups coerce and extort children into both sexual exploitation and acts of suicide and self-harm.

Animal cruelty

4.44 Alongside the potential risk of animal cruelty content being livestreamed, commenting is likely also a risk factor allowing users to conspire to commit cruelty, or to encourage or assist its commission by urging on, or suggesting or requesting specific acts, in real-time.[93]

Harms to children

4.45 There have been several cases in which one-to-many livestreaming has been used, including by children, to broadcast them self-harming or ending their life in real time.[94] This exposes children, at times unintentionally, to these livestreams and any associated comments. This creates a risk of children being exposed to comments and messages encouraging suicide and self-harm, as well as the livestream itself.[95]

4.46 The Molly Rose Foundation has stated that large volumes of comments related to suicide and self-harm risk the normalisation of self-harm as an acceptable coping strategy.[96] Other research has noted significant risk of a contagion effect, including in adolescents,[97] and

[89] Shukman, H. and Bridge, M., 2018. Paedophiles grooming children live on YouTube, The Times, 10 December. [accessed 12 June 2025].

[90] Protect Children, 2024. Tech Platforms Used by Online Child Sexual Abuse Offenders. [accessed 21 January 2025].

[91] Phillips, J G. and Mann, L., 2019. Suicide baiting in the internet era. *Computers in Human Behaviour*, 92. [accessed 4 June 2025].

[92] For more detail see Ofcom, 2024. Illegal Harms Register of Risks, p.60; National Crime Agency, 2025. Sadistic online harm groups putting people at unprecedented risk, warns the NCA. [accessed 10 June 2025].

[93] Ofcom, 2024. Illegal Harms Register of Risks, Animal cruelty chapter. [accessed 4 June 2025].

[94] We set out more details about these various cases in: Ofcom, 2025. Children's Register of Risks. [accessed 4 June 2025].

[95] Ofcom, 2025. Children's Register of Risks. [accessed 4 June 2025].

[96] Molly Rose Foundation, 2023. Preventable yet pervasive. [accessed12 June 2025].

[97] Forum on Global Violence Prevention; Board on Global Health; Institute of Medicine; National Research Council, (Gould, M., and Lake, A.), 2013. The Contagion of Suicidal Behavior. [accessed 4 June 2025].

indicates that comments mocking or encouraging victims potentially lead children to feel unable to open up a conversation about their suicidal thoughts and feelings.

4.47 In a 2023 survey of 1,423 children aged 13 to 17 in the United States, teenage gamers reported that they had experienced various forms of abuse when livestreaming gameplay. The survey found that of all 13- to 17-year-old gamers, a total of 44% (almost half) said they used Discord and 30% (almost 1 in 3) said they used Twitch, both of which involve livestreaming as part of gameplay.[98] Among the same group of gamers, 41% (over 4 in 10) of respondents said they had been called an offensive name when playing, 12% (over 1 in 10) had been physically threatened, and 8% (nearly 1 in 10) had been sent "unwanted sexually explicit things".[99]

4.48 The following evidence outlines other harms to children from comment functionality. While we have not yet identified research or evidence of these occurring on one-to-many livestreams specifically, we assess that they are very likely to have occurred. We assess that the impact of such comments when children livestream will be heightened as they will be dealing with them in the moment, with no meaningful chance to reflect and consider how or whether to respond. We invite research and evidence of these harms occurring on one-to-many livestreams in response to this consultation.

4.49 There is evidence that commenting on content has been used to encourage disordered eating behaviour, particularly among children with experience of an eating disorder. This has the potential to amplify and contribute to cumulative harm, both for those commenting and those who encounter the comments. An analysis of eating disorder content ('thinspiration' videos) found that these posts induced conversations about eating disorders. Comments were found to be a source of competition, with users comparing themselves to metrics discussed in the post and sharing how much weight they wanted to lose.[100] An investigative journalism piece reported on the trend of posting shaming comments on eating disorder content, which targeted the appearance of the user posting the content.[101]

4.50 Survey participants in a study about experiences of gaming among 8- to 17-year-olds in Australia described other gamers using hateful and abusive language (20% or 1 in 5 of respondents), expressing sexist attitudes (11% or 1 in 10) and expressing attitudes about the superiority of one race, culture, religion or nationality over another. A further 6% of respondents also said other gamers had said hurtful or nasty things to them because of "my race (or skin colour), religion, culture, nationality, disability, gender or sexuality", and 9% of 13- to 17-year-old gamers had received or asked for nude images or sexual information in comments.[102]

[98] Engagement on Twitch is likely to relate to one-to-many livestreams, while use of Discord is more likely to comprise many-to-many livestreaming between multiple players.

[99] Pew Research Centre (Gottfried, J. and Sidoti, O.), 2024. Teens and Video Games Today. [accessed 16 May 2025].

[100] Hung, M., 2022. A Content Analysis on Fitspiration and Thinspiration Posts on TikTok. *Cornell Undergraduate Research Journal*, 1(1). [accessed 4 June 2025].

[101] Polanco, J., 2020. The Incitement of Diet Culture and Disordered Eating through TikTok. J. Cole Nutrition. [accessed 12 June 2025].

[102] eSafety Commissioner, 2024. Levelling up to stay safe: Young people's experiences navigating the joys and risks of online gaming. [accessed 11 June 2025]. This research notes that of 1,799 gamers aged 8 to 17 years old who were surveyed, 66% (2 in 3) self-reported messaging or talking to people online while playing video

4.51 Other research also identified misogyny, hate, antisemitism, racism, and homophobia in comments sections when users stream and chat about games, likely in multiplayer (many-to-many) livestreaming settings.[103]

Gifting

4.52 There are multiple ways in which monetary contributions or gifts can be sent to livestreamers. One of these is the ability to send gifts 'in the moment', which can be used to influence behaviour in real time. Services often offer in-service means to make financial contributions, for example by sending tokens or gifts to other users. Payments can also be made to livestreamers via links to external gifting sites, 'wishlists', or financial payment mechanisms posted by livestreamers within their user profile or livestream.

4.53 Evidence shows that perpetrators may send money to a child or buy them virtual or physical gifts to facilitate relationship-building as part of the grooming process, to flatter children, and as a gesture to give the impression of affection to the child user.[104]

4.54 A 2022 Forbes investigative article which stated it reviewed hundreds of broadcasts on the one-to-many livestreaming service TikTok Live[105] suggests that children were broadcasting on the service, sometimes to hundreds or thousands of viewers, and adults were paying for them to perform sexual acts. Forbes reports that payments were made in the form of in-service gifting and off-service payment mechanisms listed in the children's user profiles. As well as manipulating the children by offering payment, viewers of the livestreams were reported as posting comments requesting that they pose in certain ways or show parts of their body.

4.55 The following evidence is not specific to livestreaming, but also indicates that gifting is used to groom and coerce children into sexual behaviour or other illegal or harmful activity. We include this as we consider it is almost certain to be applicable to livestreaming environments.

4.56 The 5Rights Foundation's Risky by Design research highlights that groomers may offer in-game or in-app gifts or currency to coerce them into participating in a criminal activity. Children are highly susceptible to commercial pressures, in particular relating to online games, and can easily be taken advantage of by bad actors.[106]

4.57 In response to the November 2023 Consultation, the Canadian Centre for Child Protection (C3P) outlined how some individuals who perpetrate online grooming offences send gifts

games on an online gaming service, while 26% (over 1 in 4) self-reported messaging or talking to players they didn't already know while using a gaming service. The type of livestream (one-to-many, many-to-many or one-to-one) on which these interactions took place is not specified.

[103] Miller, C. and Silva, S., 2021. Extremists using video-game chats to spread hate. BBC News, 23 September. [accessed 12 June 2025].

[104] Gámez-Guadix, M., De Santisteban, P., Wachs, S. and Wright, M., 2021. Unraveling cyber sexual abuse of minors: Psychometrics properties of the Multidimensional Online Grooming Questionnaire and prevalence by sex and age, Child Abuse and Neglect, 120. As this research notes: From a social psychology viewpoint, the principle of interpersonal influence based on reciprocity suggests that people develop a feeling of duty to return what is received from others (Cialdini and Goldstein, 2004). Thus, the use of gifts in online grooming situations can predispose minors to acquiesce the sexual requests made by the adult.

[105] Levine, A., 2022. How TikTok Live Became 'A Strip Club Filled With 15-Year-Olds', Forbes, 27 April. [accessed 10 June 2025].

[106] 5Rights Foundation, 2020. Risky by Design. [accessed 4 June 2025].

and tokens to children as part of the grooming process or to incentivise the child to share sexual imagery.[107]

Reaction features

4.58 'Reaction features' comprise the use of a user-to-user service functionality which allows users to express a reaction, such as approval or disapproval, of content that is shared by other users, through dedicated features that can be clicked or tapped by users. This for instance includes 'liking' or 'disliking' a piece of content.[108]

4.59 As noted by the 5Rights Foundation, design choices such as hearts that visualise 'likes' enable others to exploit the desire for social affirmation, which is strong in children and young people. 5Rights state that children as young as seven to 10 years old have been pressured into performing sexual acts on livestreams in exchange for likes.[109]

4.60 The IWF carried out a study in 2017 to examine characteristics of captures of livestreamed CSAM. Among the 2,082 images assessed, it was apparent in several cases that children had been manipulated into sexual activity on livestreams to gain 'likes' from viewers. The study also noted that perpetrators may capitalise on children's desire for belonging, 'rewards' or celebrity-like status by 'gamifying' the grooming process using 'likes'. For example, a child may initially agree to sexual posing on receiving 500 'likes' and may then agree to further acts as the 'game' proceeds.[110]

4.61 Analysis of over 39 million chat messages exchanged during one-to-many livestream broadcasts on Live.Me found that perpetrators used emojis as a form of communication to groom children into performing sexually inappropriate and suggestive acts, such as requesting the removal of clothes.[111]

4.62 While not specific to livestreaming, the following evidence indicates how children can be vulnerable to behaving in ways contrary to their wellbeing or safety in return for 'likes'. We assess that this evidence is likely to be applicable to livestreaming, where the immediacy of the medium places children at heightened risk of harm through impulsive actions.

4.63 Gaining engagement from others on the content they post is an important motivating factor for children to participate in harmful and non-harmful challenges. A survey of 5,400 teenagers aged 13 to 19 in several countries, including the UK, found that the most common reason teenagers took part in challenges was to gain views, comments and 'likes'.[112]

[107] Canadian Centre for Child Protection response to November 2023 Illegal Harms Consultation, p.2.

[108] As set out in the glossary of: Ofcom, 2024. Illegal Harms Register of Risks. [accessed 12 June 2025]. This definition may include the following, as noted in section 4(1)(c) of the Online Safety Act 2023: (i) applying a "like" or "dislike" button or other button of that nature, (ii) applying an emoji or symbol of any kind, (iii) engaging in yes/no voting, or (iv) rating or scoring the content (or the comments or reviews) in any way (including giving star or numerical ratings).

[109] 5 Rights Foundation, 2025. Case Study 2: Livestreaming and video sharing. [accessed 4 June 2025].

[110] Internet Watch Foundation, 2018. Trends in Online Child Sexual Exploitation: Examining the Distribution of Captures of Live-streamed Child Sexual Abuse. [accessed 4 June 2025].

[111] These chat messages were exchanged by more than 1.4 million users in 291,000 live broadcasts over two years: Lykousas, N. and Patsakis, C. 2021. Large-scale analysis of grooming in modern social networks. Expert systems with applications. [accessed 4 June 2025].

[112] Hilton, Zoe. Praesidio Safeguarding, 2021. Exploring effective prevention education responses to dangerous online challenges. [accessed 10 June 2025].

4.64 The psychological impact that 'likes' can deliver is likely to encourage children to increase their level of use. Functionalities such as receiving likes are designed to be exciting; the 'likes' pour in as a form of social affirmation. The dopamine 'hit' of getting a 'like' strongly encourages the young person to continue to behave in the ways that are most 'liked'. [113]

4.65 Validation from other users on a service, including through 'likes', comments or re-posting, can reinforce or even exacerbate negative thought patterns or behaviours, and potentially encourage the further posting of potentially harmful content. It can also provide users with a sense of community in feeling that they are not alone in their thinking.[114]

Content capture and screen recording

4.66 Some livestreaming services have dedicated in-service screen recording and screen capture functionalities, which viewers can use to record videos or still images of livestreams. Users can also screen-record or capture content using third-party services. In the latter case, the livestreamer is particularly unlikely to be aware that their content is being recorded.

CSEA content

4.67 There is evidence to indicate that screen recording and screen capturing on livestreams forms the majority of CSAM that is generated. The IWF's 2021 annual report highlighted that just over 70% (almost three quarters) of reports of CSAM in that year involved recordings taken via a phone or webcam of children, who were often alone in their bedrooms.[115] These images and videos had often been derived from livestreaming services. Over half (59%) of such reports specifically showed the sexual abuse of 11- to 13-year-old girls who had been groomed, coerced, or encouraged into sexual activities via a webcam.[116] While we do not know the amount or proportion of CSAM that was derived from one-to-many livestreams over other livestreams, we consider it likely to be sizable.

4.68 A 2024 investigation by C3P and Bloomberg examined Twitch's 'Clips' archive of short videos up to 20 seconds long. These clips are made by viewers of their favourite (likely one-to-many) livestream moments. The investigation found that 83 (7.5%) of a sample of 1,100 clips could be classified as featuring sexually explicit depictions of minors. The 83 clips were viewed around 10,000 times.[117] The study indicated that, in most instances, the children had been groomed into this behaviour.

4.69 We assess that comment, gifting, and reaction functionalities are likely to be used by perpetrators to achieve CSAM capture. Such content then proliferates online and is shared by perpetrators, which places it outside of the victims' control and puts them at greater risk of further coercion and sexual abuse. For example, explicit images or videos of a child

[113] 5Rights, 2023. Disrupted Childhood: The cost of persuasive design. [accessed 10 June 2025].

[114] Biddle, L., Derges, J., Goldsmith, C., Donovan, J L. and Gunnell, D., 2018. Using the internet for suicide-related purposes: Contrasting findings from young people in the community and self-harm patients admitted to hospital, p.12, PLoS ONE, 13 (5). [accessed 10 June 2025].

[115] Internet Watch Foundation, 2021. IWF Annual Report. [accessed 15 January 2025]. In 2021, the IWF assessed 361,062 reports, of which 70% (252,194, or nearly 3 in 4) led to the identification of imagery of children online being sexually abused.

[116] The IWF has not undertaken granular analysis of the provenance of CSAM in relation to livestreaming in subsequent years to provide a more up-to-date assessment. However, they indicated in January 2025 that this remains a prevalent issue.

[117] Bloomberg UK, 2024. Twitch 'Clips' Feature Is Used by Predators to Record and Share Child Abuse. [accessed 11 June 2025]; Winslow, L., 2024. Report: Predators Are Using Twitch Clips To Spread Child Abuse, Kotaku, 5 January. [accessed 4 June 2025].

can be used to blackmail them into sharing further images or to incite the child to abuse friends and siblings. Perpetrators may do this by threatening to publish the content online, send the images to friends and family,[118] or through financially motivated sexual extortion.[119]

4.70 For victims and survivors, knowing that CSAM remains available online and that perpetrators may still be using it can be a continuing source of trauma. Some describe achieving a sense of 'closure' as impossible.[120] Many describe feeling constantly fearful and vulnerable because their abuse exists as a permanent record online which others can view.[121]

4.71 The negative impacts on the child depicted can include mental health challenges[122] and negative social consequences,[123] particularly for girls, who will often face bullying, harassment, social exclusion, and victim-blaming.[124] Many victims and survivors describe feelings of self-blame, negative psychological health, and heightened anxiety from knowing that the images remain online.

Terrorism content

4.72 Content capture is also a risk factor for livestreaming terrorism content. Although a terrorist attack may only be live for a short period, there is evidence that the content can be recorded and re-shared to reach a significantly larger audience.[125]

4.73 The spread of such content has impacts offline, including potentially inspiring further far-right attacks. For example, the attack in May 2022 in Buffalo, New York was livestreamed online, and versions of the footage were disseminated on multiple online services[126].

4.74 The video of the Christchurch mosque attacks in New Zealand, in March 2019, was watched approximately 4,000 times before it was removed. Footage of the attack was then shared on other sites; in the 24 hours that followed the attack, Facebook removed 1.5 million videos of the attack and a further 1.2 million were blocked at upload.[127]

[118] National Crime Agency, 2024. Child sexual abuse and exploitation. [accessed 10 June 2025].

[119] NSPCC, 2024. Young people's experiences of online sexual extortion or 'sextortion'. [accessed 10 June 2025].

[120] CSA Centre (Brown, S.), 2023. Key messages from research on child sexual abuse by adults in online contexts. [accessed 4 June 2025].

[121] Owens, J. N., Eakin, J. D., Hoffer, T., Muirhead, Y., Lynn, J., and Shelton, E., 2016. Investigative aspects of crossover offending from a sample of FBI online child sexual exploitation cases. Aggression and Violent Behaviour, 30. [accessed 4 June 2025].

[122] Frankel, A., Bass, S., Patterson, F., Dai, T. and Brown, D., (2018). Sexting, Risk Behaviour, and Mental Health in Adolescents: An Examination of 2015 Pennsylvania Youth Risk Behavior Survey Data. Journal of School Health, 88(3). [accessed 4 June 2025].

[123] 3 From a qualitative study with 41 young people in south-east England. Setty, E. 2019. A rights based approach to Youth Sexting: Challenging, Risk, Shame, and the Denial of Rights to Bodily and Sexual Expression Within Youth Digital Sexual Culture. International Journal of Bullying Prevention, 1. [accessed 4 June 2025].

[124] Ringrose, J., Regehr, K., Whitehead, S., 2022. 'Wanna Trade?' Cisheteronormative homosocial masculinity and the normalisation of abuse in youth digital sexual image exchange. Journal of Gender Studies, 31(2). [accessed 4 June 2025].

[125] Ofcom, 2024. Illegal Harms Register of Risks. [accessed 4 June 2025].

[126] Ofcom, 2022. The Buffalo Attack: Implications for Online Safety. [accessed 4 June 2025].

[127] Ofcom, 2022. The Buffalo Attack: Implications for Online Safety. [accessed 4 June 2025].

Hate content

4.75 Screen capture can contribute to the risk from hateful content. The subsequent recording and dissemination of potentially hateful livestreamed footage can increase the proliferation of the content online.[128]

Our approach to livestreaming measures

4.76 Based on the detail discussed in this section, we are now consulting on measures that will go further to protect users from the risks associated with livestreaming. We have taken a two-pronged approach to doing so.

4.77 In chapter five we set out our approach to protecting all users from encountering illegal harms (including terrorism content, CSAM and grooming, incitement to violence, content assisting or encouraging suicide, hate and animal cruelty).

4.78 In chapter six we set out our proposals to protect child users in the UK when they livestream. We are proposing to introduce these as new measures within our Illegal Harms U2U Codes.

4.79 In chapter seven we detail our position on further steps we may consider taking in relation to livestreaming.

[128] Ofcom, 2024. Illegal Harms Register of Risks. [accessed 4 June 2025].

5. Livestreaming – our proposals to protect all users

Our proposals

Number	Proposed measure	Who should implement this
ICU D17	The provider should have a mechanism to enable users to report that a livestream contains content that depicts the risk of imminent physical harm**.**	All user-to-user services offering one-to-many livestreaming at medium or high risk of terrorism content, grooming or image-based CSAM, assisting or encouraging suicide, hate, or harassment, stalking, threats and abuse offences. This includes services that only provide livestreaming and those where it is an ancillary feature.
ICU C16	The provider should, as part of its content moderation function (see ICU C1), ensure that human moderators are available whenever users can livestream using the service	All user-to-user services offering one-to-many livestreaming at medium or high risk of terrorism content, grooming or image-based CSAM, assisting or encouraging suicide, hate, animal cruelty, or harassment, stalking, threats and abuse offences. This includes services that only provide livestreaming and those where it is an ancillary feature.

Consultation questions

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

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

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

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

Our proposals

Our provisional view is that a minimum set of measures is required to set a baseline to protect users

5.1 We are proposing measures to help strengthen user reporting and content moderation for services with a livestreaming functionality, regardless of service size. This will mean that such services will be able to better and more quickly identify, prioritise and take timely action against suspected illegal content, in particular for harms that are detailed in our Risk Profiles for livestreaming services.[129]

5.2 These measures will especially help protect against illegal activity where there is a risk of physical harm.[130] This content can generate significant harm for viewers, members of the public or, in some cases, livestreamers themselves.

5.3 These measures mainly concern incidents where there is a livestreamed terrorist or hateful attack against protected groups with the potential of mass casualties, real-time child sexual abuse, a livestreamer considering or attempting suicide where viewers might be encouraging them, and incitement of hate or violence (including during a crisis).

5.4 There are existing measures in our Illegal Content User-to-User Codes relating to reporting and complaints and content moderation, as described in paragraphs 5.7-5.12 below. Our proposals build on these measures for services with a livestreaming functionality.

5.5 We are proposing two specific additional measures:

- ICU D17 requires that livestreaming services have a specific reporting category that allows a user report to specify whether the livestream depicts the risk of imminent physical harm. Our policy intent is that this would include a threat to life and CSEA.
- ICU C16 requires that livestreaming services should, as part of their content moderation function, ensure that human moderators are available whenever users can livestream using the service.

Interaction with the regulatory framework

5.6 The proposed measures complement our existing reporting and content moderation measures in our Illegal Content User-to-User Codes.

Reporting and complaints measure

5.7 Our Illegal Content User-to-User Codes[131] recommend that providers should design and operate complaints procedures so that they are easy to find, access, and use (ICU D2). This includes complaints about potentially illegal content.

[129] For illegal content this includes CSEA, terrorism, suicide, hate, harassment/stalking/abuse, and animal cruelty content. Our Protection of Children Risk Profiles note that livestreaming is a risk factor for suicide and self-harm and violent content.

[130] This could include child sexual abuse, a live terrorist attack, or content relating to a potential suicidal attempt.

[131] Ofcom,2024, Illegal content Codes of Practice for user-to-user services [accessed 11 June 2025]

Content moderation measures

5.8 Content moderation measures in our Illegal Content User-to-User Codes[132] recommend that all user-to-user providers should, as part of their content moderation function, have systems and processes designed to review and assess content that the provider has reason to suspect may be illegal content (ICU C1).

5.9 When a provider has reason to suspect that content may be illegal content, the provider should review the content and either make an illegal content judgement or if its terms of service prohibit the type of illegal content, should consider whether it is a breach of their terms of service (illegal content proxy).

5.10 Providers should, as part of their content moderation function, have systems and processes designed to swiftly take down illegal content and illegal content proxy of which they are aware unless it is currently not technically feasible for them to achieve this outcome (ICU C2).

5.11 Large and/or multi-risk services should set performance targets, and in doing so they should balance speed and accuracy (ICU C4). This means that in-scope providers should balance the need to take content moderation action swiftly, against the importance of making accurate content moderation decisions.

5.12 Large and/or multi-risk services should resource their content moderation function to give effect to their performance targets (ICU C6). In doing so, they should have regard to the propensity for external events to lead to a significant increase in demand for content moderation.

Explanation of our proposed measures

The provider should have a mechanism to enable users to report that a livestream contains content that depicts the risk of imminent physical harm

5.13 User reporting is an important element for alerting services to illegal content on livestreams as a supplement to proactive tools.

5.14 We consider it particularly important that there is a mechanism allowing users to report a livestream that depicts the risk of imminent physical harm. This means that services can be provided with a specific signpost to those livestreams that may contain this kind of severely harmful content.

5.15 Under this measure, services should ensure that when UK users report content within a livestream, they should be able to clearly indicate that the content they are reporting depicts the risk of imminent physical harm. This could include child sexual abuse, a live terrorist attack, or content relating to a potential suicidal attempt.

5.16 We consider that categories such as ‘dangerous activities and challenges’, ‘violent content’, ‘child safety’, ‘terrorism content’ or ‘suicide and self-harm’ would not comply with this measure because the categories are not specific to the risk of imminent physical harm.

[132] Ofcom,2024, Illegal content Codes of Practice for user-to-user services [accessed 11 June 2025]

5.17 If a service already has a category that allows user to report that the livestream shows imminent physical harm, then the service is already following this measure.

5.18 Services should ensure that they have other categories for reporting content that might be illegal, but where there is no imminent risk for physical harm (for example, terrorism, fraud or hateful content). This will help ensure this reporting category is used only for livestreams where the risk of harm is the most severe.

5.19 This measure will mean that suitable assessments can be made by services on the livestreamed content. Services will be more likely to stop or take down the livestream midway, limiting the harm to users and the public.

5.20 This measure applies to UK users. However, services may also choose to give users in other jurisdictions the ability to report in this way.

The provider should ensure that human moderators are available whenever users can livestream

5.21 Given the limitations of proactive technology in the livestreaming context (as set out below) we are proposing that services have human moderators available at any time that the livestreaming functionality is available to users. This is to ensure that content that is suspected to be illegal can be reviewed and assessed as close to the time it is flagged as possible.

5.22 The presence of human moderators will supplement automated tools whose effectiveness on livestreams varies and is dependent on the harm and the type of content. To comply with this measure, human moderators should be non-volunteer moderators.

5.23 Automated content classifier tools are more able to detect content such as nudity rather than live violent or graphic content such as a terrorist or hateful attack or a suicide or attempted suicide. In addition, as some classifiers analyse content at intervals, illegal content could be missed, leading to the livestream remaining online after an incident is over.

5.24 The use of automated tools to detect harms such as Child Sexual Exploitation and Abuse (CSEA) in livestreaming environments present several challenges. These include the complexity of analysing live video versus still images or pre-recorded content, the difficulty of identifying real-time instances of child sexual abuse rather than known child sexual abuse material, and strict regulatory and ethical constraints governing the training of detection models. These factors often necessitate the involvement of human moderators.

Limitations of proactive tools

Used alone, proactive tools have some limitations for detecting certain types of harm in livestreamed content.

The effectiveness of automated content moderation tools depends on their underlying technical architecture and data, with key performance metrics including precision, recall, processing speed, language coverage and contextual understanding.

Broadly, automated content moderation tools need to overcome several points of complexity in a livestreaming context.

These include:

Latency constraints: where the speed of detection and action must be balanced against accuracy through false positives and false negatives.

Multi-modal complexity: the need to respond to violations that may span video, audio and chat simultaneously requiring multiple automated content moderation systems used in combination.

Cost considerations: the continuous data flow and real-time processing demands carry inherent resource implications.

Given these factors, each service might apply different tolerance thresholds for each of the Automated Content Moderators (ACMs) used depending on the size of their service and their human moderation capacity.

5.25 We note that as part of the existing content moderation resourcing measure (ICU C6) large/multi-risk services may already have human moderation. However, ICU C6 does not explicitly recommend that human moderators be available when a user can livestream so we consider that this measure is required as a backstop for livestreaming services.

Benefits and effectiveness at addressing risks

Reporting category

5.26 Services have discretion on the categories and design of reporting systems, provided that they include a reporting category for imminent risk of physical harm, which includes threats to life and CSEA.

5.27 From a user perspective, we are recommending a single additional category that will enable reports of the risk of imminent physical harm on livestreaming. While we do not consider this additional category as adding to user-choice burdens, we recognise there are limits to how many categories a user can decide between and do so accurately.[133] Therefore, services have discretion over the categories and design of reporting systems, and we welcome any evidence to help us understand these further.

5.28 We considered whether an additional option when in the reporting flow would confuse users and make them less likely to report. However, our research suggests that users do not struggle with submitting a report once they are in the reporting flow. We therefore conclude that users will select an accurate category.[134]

[133] Chernev, A., Böckenholt, U. and Goodman, J., 2015. Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, *25*(2), pp.333-358 [accessed 11 June 2025]. Choice overload is an important concept from behavioural science. A greater number of choices makes things cognitively complex for people and mean they make worse decisions.

[134] Ofcom, 2023. Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour. [accessed 11 June 2025]. Within a research setting using legal but potentially harmful content, we did not find evidence to suggest changing reporting categories impacted the likelihood or accuracy of users reporting. Given the other findings of the trial, it suggests that once into the reporting flow, users did not struggle with submitting a report.

5.29 User reports are important for flagging harmful content to services. Almost 1 in 5 (18%) of UK users said they used the function or similar when they encountered an online harm.[135] We do not consider that this measure will increase the overall number of reports a service receives. Therefore, this measure is unlikely to place a high burden on services.

5.30 Expert opinion recommends that online services have a way for people to report suicide content[136] and suggests that reporting potentially illegal suicide content is crucial to keeping users safe.[137]

Human content moderation

5.31 Human moderation being available whenever users can livestream will mean illegal content will be removed more quickly because there will be someone available to make effective determinations on content that is suspected as illegal.

5.32 We have evidence that taking quick action on livestreamed terror attacks can minimise their impact. In the 2019 Christchurch attack, footage remained online for the duration of the attack and was recorded and shared widely. In the 24 hours that followed, Facebook removed 1.5 million videos of the attack and a further 1.2 million were blocked at upload. A year after the attack, researchers at the Counter Extremism Project identified at least 14 different websites where footage of the Christchurch terror attack could be accessed.[138]

5.33 In contrast, the Buffalo attack in 2022[139] was stopped by the service midway through the livestream. There were 28 viewers, or fewer, who watched the livestream of the channel at some point during broadcast. Despite this low figure, footage of the attack was also spread across platforms and seen by millions of people. However, the footage had less virality and longevity online than the Christchurch footage as the livestreaming was stopped midway due to moderation actions from services[140].

5.34 The online dissemination of the Buffalo attack was in part the result of a small service called Streamable which acted slowly to remove a video of the livestream. In addition, we are aware that some services did not have round-the-clock moderation in place at the time of the Buffalo attack.[141] [142]

5.35 Taking quick action on content relating to a potential suicide attempt, including that which is livestreamed, can minimise its impact as it will prevent users from encouraging this act via comments. It also has the potential of disrupting a person from attempting suicide and reducing the number of viewers seeing it. Samaritans' guidelines for service providers on

[135] Ofcom, 2025. Experiences of using Online Services - Wave 7 [accessed 11 June 2025]. Of all respondents who have experienced an online harm in the past four weeks and know their most recent experience, 25% (1 in 4) said "I scrolled past it" and 18% (almost 1 in 5) said "Clicked the report/flag button or marked as junk".
[136] Samaritans, 2020. Managing self-harm and suicide content online [accessed 8 May 2025].
[137] Samaritans. What to do if you see worrying suicide and self-harm content online [accessed 8 May 2025].
[138] Office of the New York State, 2022. Buffalo Shooting Online Platform Investigative Report [accessed 12 June 2025]
[139] On 14 May 2022, an attack was carried out in Buffalo, New York, which resulted in the death of ten individuals and the wounding of three others. The attack was livestreamed online and versions of the footage were disseminated on multiple online services, potentially exposing UK users to content related to terrorism.
[140] Office of the New York State, 2022. Buffalo Shooting Online Platform Investigative Report [accessed 12 June 2025]
[141] Ofcom, 2023. BitChute improves its safety measures following engagement with Ofcom - Ofcom [accessed 12 June 2025].
[142] [✂.]

managing self-harm and suicide content online note that for providers to protect users from potentially harmful suicide content, they need to have effective processes for reviewing and responding to user reports, where those relating to users in need of urgent help should be prioritised for review.[143]

Overall effectiveness

5.36 Together these measures should increase the likelihood that a livestream depicting a violent attack or CSEA is removed faster, reducing the number of viewers and decreasing the risk of harm. Taking action on live suicide content may also potentially stop the user from going through with an attempt at suicide if, for example, viewers are prevented from encouraging the act, for instance via comments.

5.37 Furthermore, these measures are likely to reduce the value of any recordings of the livestream that may be subsequently reloaded and shared on other services. This is because incomplete footage is less appealing and is less likely to be circulated as widely (as set out in paragraphs 5.32 ad 5.33 above) meaning a decreased impact in terms of, for example, radicalisation.

5.38 There may be instances where these measures would deter would-be users from carrying out the act of physical harm. This would apply to circumstances where reaching a wide audience is a key motivating factor in the act. There is evidence from previous livestreamed terror attacks that livestreaming the event is central to the attackers' motivation. For example, prior to the 2022 livestreamed terrorist attack in Buffalo, the attacker talked about gaining infamy by following in the footsteps of others who had livestreamed hateful attacks. He also considered audience reach and impact when choosing the service he used for the livestream.[144]

5.39 Together, we expect improved reporting and better human moderation coverage could mean terrorist, CSEA and suicide livestreams are disrupted and removed more quickly. This would potentially lead to the following safety outcomes:

- reduce harm to viewers;
- reduce radicalisation towards terrorism and encouragement of suicide;
- lead to incomplete recording of the content, which would have less sensationalised impact for those wishing to disseminate copies of such footage;
- reduce the risk of harm from comments on livestreams (this is especially relevant to abuse, hate, and encouragement of suicide); and
- reduce the reach and impact of livestreamed terror attacks, and therefore the potential for attackers to gain notoriety. This in turn could deter would-be attackers (where the ability to broadcast real-world harm in real time is part of the plan of carrying out the act).[145]

[143] Samaritans, 2020. Managing self-harm and suicide content online [accessed 8 May 2025].

[144] Office of the New York State, 2022. Buffalo Shooting Online Platform Investigative Report (pg 31) [accessed 12 June 2025]

[145] Office of the New York State, 2022. Buffalo Shooting Online Platform Investigative Report [accessed 12 June 2025]. There is evidence that the Buffalo attacker's perception of the efficacy of different services' content moderation was a factor in selecting which service he decided to use.

Technical feasibility

5.40 We assess that the addition of 'imminent physical harm' as a reporting category option will be technically feasible for services to implement, as well as any changes required to material referencing said functionality. For more detail on implications for surrounding systems and processes, and prioritisation of reports made using this category, refer to Volume 2, Chapter 2, Protecting people from illegal harms online.

Impacts and Costs

Direct financial costs for service providers

Reporting category

5.41 Setting up a reporting category will require services to incur a one-off implementation cost and ongoing maintenance costs. Implementation costs will include designing required changes, engineering costs of testing and implementing those changes, and the costs of further updates and refinement. We estimate that it could take two to three-and-a-half months of full-time work for a software engineer to implement a reporting category, with a similar time input from other skilled professionals such as product managers, analysts and lawyers. This results in a one-off implementation cost between £19,000 and £68,000. These costs will likely vary depending on the complexity of the service (smaller services may be less complex in terms of technical set-up and governance), as well as the existing set-up of reporting on the service. Ongoing maintenance costs will primarily consist of labour costs, and we estimate the annual costs of these to range between £5,000 and £17,000.[146]

Human content moderation

5.42 The main cost driver of the content moderation measure is the additional labour cost of employing or contracting more human moderators. This will be an ongoing cost that may adjust over time to accommodate changes in service demand and use. Costs will scale depending on the number of human moderators required. In general, we expect this to be greater for larger services with higher volumes of content.

5.43 To be effective, human moderators may require specific tools or training to detect and review content. They may also require an ICT support team. Service providers may also decide to offer mental health support and other wellbeing benefits to their content moderators, which would add to costs. We have considered the inclusion of outsourced content moderation in this measure and note that content moderation costs could be lower where this is the case.

5.44 Hypothetically the costs may be greater for small and medium-sized single risk services who will not be in scope of ICU C6 (covering resourcing for content moderation). We are not currently aware of any livestreaming services that would fall in this category, though a small number of such services may exist.

5.45 Providers that already have policies and processes in place that are sufficient to implement this measure would not need to incur any additional costs.

[146] This is based on our cost assumptions outlined in Annex 15.

Wider market impacts

5.46 Imposing additional costs of establishing and running a livestreaming service will have a larger relative impact on smaller services and could potentially dissuade market entry, risking market consolidation. There is a risk that some service providers may exit or choose not to enter the market as a result of our proposed measures. This may be more likely for services where livestreaming is a secondary or lower priority part of their service provision. However, based on available data, we expect the actual risk of market exit to be low – most services that have a livestreaming functionality (either as a primary or additional function) are medium or large services and we consider that establishing a livestreaming functionality or service already has high fixed costs.

Rights assessment

Freedom of expression and association

5.47 In Chapter 3 we set out the detail of the rights to freedom of expression and association, including that we must exercise our functions under the Act in a way that does not restrict these rights unless we are satisfied that it is necessary and proportionate to do so.

Reporting category

5.48 We consider that this measure does not directly interfere with users' rights to freedom of expression and association.

5.49 In Chapter 14 we are recommending that services exclude from recommender systems content potentially indicated to be certain kinds of priority illegal content, which may include livestreams reported under this category. Therefore, a potential indirect restriction of rights arises where a service, as a result of a report labelled in this way, removes the livestream from recommender feeds in accordance with our recommender systems measure. This interference is particularly significant where the reported livestream is not in fact illegal content, as removal from recommender feeds may negatively impact real-time engagement with the livestream.

5.50 However, if a service determines that the livestream is not illegal content (or an illegal content proxy), it can reinstate the livestream in recommender feeds (or, if the livestream has ended, a video of it). To this extent, the interference with rights would be temporary. Given the seriousness of potential harm arising from livestreams (as set out in chapter 4) we do not consider the interference to be disproportionate.

5.51 We also appreciate that when a provider moderates reports labelled 'imminent physical harm' they may take an overly cautious approach by taking down content which is not illegal (or an illegal content proxy), presenting an impact on the users right to freedom of expression. However, our Illegal Content Judgements Guidance is intended to assist services in making this judgment based on reasonably available information. Providers should consider reports as part of deciding what content moderation action to take. In addition, other measures ensure that users have a mechanism to appeal incorrect removal decisions (and we are proposing to broaden our appeals measures to ensure they enable appeals of decisions that content was illegal content proxy).

5.52 Where livestreams reported under this measure are indeed illegal content (or an illegal content proxy), and are therefore taken down, this reflects the duty for services to take down illegal content pursuant to the requirements of the Act. We do not consider this to constitute a disproportionate interference with freedom of expression.

5.53 This removal will limit users' exposure to such livestreams (including through onward sharing or copying). This could also have positive impacts on freedom of expression where users feel more confident engaging with livestreams if they know that content of this nature can be reported easily and that they are less likely to encounter it.

5.54 Given the limited direct interference posed by the measure, and the safeguards to mitigate the indirect interferences referred to in paragraph 5.51, we consider that the potential interferences are proportionate having regard to the benefits that the measure provides in assisting providers to take appropriate moderation action in relation to illegal content.

Content moderation

5.55 As with any content moderation decision, including human review, there is a risk of incorrect assessment and removal of content and providers taking an overly cautious approach to moderation. There will be an interference with users' freedom of expression where the livestream is not illegal content.

5.56 However, we consider this risk is mitigated by a number of other measures, including:

a) ICU C7: large and multi-risk service providers should ensure that individuals working in content moderation receive training and materials enabling them to carry out their role.

b) ICU D8, D9 and D10: service providers should determine appeals of content moderation decisions promptly. When deciding to reverse a decision that content was illegal, they should reverse the action taken (as far as appropriate and possible).

5.57 It is also mitigated by the fact that we have provided guidance in our Illegal Content Judgements Guidance to assist providers in making moderation decisions. Furthermore, providers have incentives to limit incorrect moderation decisions, to meet users' expectations and avoid the costs of dealing with appeals.

5.58 Content moderation may also have a positive impact on freedom of expression where users may feel more confident to engage with livestreams if they know they are less likely to encounter illegal content on the service due to effective content moderation.

5.59 Our view is that human moderation is the most effective tool available to ensure services can make timely moderation judgments on livestreamed content that has been flagged. Therefore this measure, which recommends that human moderators are available whenever users can livestream using the service rather than relying solely on automated moderation decisions is likely to result in improved quality of moderation decisions. This could reduce the number of incorrect removals of legal content and could therefore have a positive impact on users' freedom of expression.

5.60 Overall, we consider the interference to freedom of expression rights as a result of this measure is limited and likely to be proportionate.

Privacy and data protection

5.61 Article 8 of the European Convention on Human Rights (ECHR) sets out the right to respect an individual's private and family life.

Reporting category

5.62 Dealing with reports inevitably involves processing personal data. Reports made under this category may be human reviewed as part of the moderation process. This may impact users' rights to privacy and their rights under data protection law. However, we consider that the impact on users' privacy rights will not be disproportionately impacted as service providers are required to comply with relevant data protection legislation when processing personal data.

5.63 The moderation may also impact the privacy rights of people depicted in the livestream. However, this is to be balanced against the greater interference with those individuals' privacy rights if the content were to remain on the service unmoderated.

5.64 We also consider this interference proportionate given the very significant benefits of content moderation in reducing harm.

Content moderation

5.65 This measure presents similar considerations in relation to privacy rights as the reporting category measure set out in paragraphs 5.62 and 5.63.

5.66 However, we consider this impact on privacy rights to be proportionate in light of the clear benefits of content moderation in identifying and appropriately actioning illegal content.

5.67 To the extent that moderation involves the processing of personal data, service providers are required to comply with the relevant data protection legislation.

Which providers should implement this measure

Reporting category

5.68 We propose that this measure would apply to all user-to-user one-to-many livestreaming services that are medium or high risk of one or more of the following harms: terrorism; grooming or image-based CSAM; assisting or encouraging suicide; hate; or harassment, stalking, threats, and abuse offences.

5.69 As we have shown, livestreaming functionality often plays a significant role in disseminating content related to these harms. Our analysis suggests that enabling users to report imminent physical harm to humans would play a meaningful role in reducing this content on livestreaming services. Given the severity and prevalence of the harm and the relatively low cost of implementing the proposed measure, we provisionally conclude that it would be proportionate to recommend this measure to all services with livestreaming functionalities that are risky for these harms.

Human content moderation

5.70 We propose that this measure would apply to all user-to-user one-to-many livestreaming services that are medium or high risk of one or more of the following harms: terrorism; grooming or image-based CSAM; assisting or encouraging suicide; hate; animal cruelty; or harassment, stalking, threats, and abuse offences.

5.71 Under existing measures in the Illegal Content Codes, if a user-to-user service is large and/or multi-risk, including those with a livestreaming functionality, it will need to ensure that the moderation it deploys is appropriately resourced to meet its performance targets.[147] However, due to the significant risks of harm being disseminated on livestreaming services and the challenges of moderating livestreams using other methods, we consider it proportionate to strengthen the Codes by ensuring livestreaming services have human moderation available whenever users can livestream.

5.72 We consider there is significant risk that illegal content could be diverted to smaller services where moderation is currently weaker if only large services were in scope of this measure. There is considerable evidence that priority illegal content, including terrorism and CSEA is broadcast and captured from small livestreaming services as well as large services, and that content encouraging or assisting suicide as part of a livestream also appears on large and small services. There is substantial benefit to having effective moderation on all services, regardless of size, to mitigate harms and reduce the effect of displacement. We therefore provisionally conclude that this measure should apply to all providers of one-to-many livestreaming services, including small services.

Provisional conclusion

5.73 The proposed measures detailed in the chapter will result in more illegal livestreamed content being detected and removed by services while the livestream is still in progress.

5.74 The measures ensure that users have an option to report on the most imminent harms and in addition, the scale and severity of risk from livestreamed illegal content is matched by having human moderation in place when livestreaming is available.

5.75 UK users should be always afforded minimum protections from content. We are aware of the cost implications, particularly for the content moderation measure, however, we consider that the costs are proportionate given the risk that livestream content can pose to UK users.

5.76 Having considered the cost implications on all services, including smaller services offering livestreaming functionality, we consider that these measures are both effective and proportionate to balance the heightened risk to users.

[147] ICU C4, ICU C6.

6. Livestreaming – our proposals to protect children

Warning: this chapter contains content that may be upsetting or distressing

Our proposals

Number	Proposed measure who	Who should implement this
ICU F3	Providers should ensure that users are unable to do the following in relation to a one-to-many livestream by a child in the UK: a) Comment on the content of the livestream; b) Gift to the user broadcasting the livestream; c) React to the livestream; d) Use the service to screen capture or record the livestream; e) Where technically feasible, use other tools outside of the service to screen capture or record the livestream.	Services that: a) offer one-to-many livestreaming, and b) it is possible for children to access the service, or a part of it.

Consultation questions

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

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

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

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

Our proposals

Our view is that a minimum set of measures is required to set a baseline to protect children

6.1 Given the scale and severity of the harms outlined in chapter four, ‘Livestreaming – Introduction’, we consider that action is needed to increase protections for children when they livestream. We are therefore proposing the inclusion of a further measure in the

Illegal Content User-to-user (U2U) Codes to make children’s experience of livestreaming safer across all services.

6.2 Our proposals are focussed on addressing the risks to children from undertaking one-to-many livestreams as described in chapter four, as we assess that there is a strong case to go further than we currently do in the Codes of Practice to mitigate these risks.

6.3 We already have in place measures to protect against the viewing of livestreamed illegal content and content harmful to children in the Illegal Content U2U Codes and the Protection of Children U2U Codes, such as our content moderation and complaints measures. We note however that it is technically challenging for services to deploy scalable technology solutions to moderate livestreams in real time. We will continue to monitor how effective our Codes are at reducing exposure to illegal content and content harmful to children when viewing livestreams, and whether additional measures are required. We also anticipate that our proposals will reduce children’s exposure to such content by helping to reduce its availability.

6.4 The proposals we are putting forward target functionalities that, particularly because of the ‘in the moment’ nature of livestreaming, exacerbate the risk of harm to children. Further, as stated in our Illegal Harms Register of Risks and Children’s Register of Risks, commenting, gifting and reaction functionalities pose risks in themselves and in association with other services, such as social media. All services likely to be accessed by children need to consider how to mitigate these risks as part of their protection of children duties. We will continue to monitor the risk of harm inherent to these functionalities and how services are managing the risks to children to evaluate whether additional measures are required.

Interaction with the regulatory framework

6.5 Under section 10(2) of the Online Safety Act 2023 (the Act) user-to-user services have a duty to prevent individuals from encountering priority illegal content, to effectively mitigate and manage the risk of the service being used for the commission or facilitation of a priority offence, and to effectively mitigate and manage the risks of harm to individuals.

6.6 To the extent that we are recommending a measure in the Illegal Content U2U Codes, we are recommending that providers use highly effective age assurance to determine which users are adults and therefore apply the measure appropriately. We set out proposals for cross-cutting measures for the implementation of highly effective age assurance in chapter 18.

Explanation of our proposed measures

6.7 We propose that services implement all of the following for one-to-many livestream broadcasts:

- When children in the UK undertake a one-to-many livestream broadcast, viewers should only be able to watch and should not be able to interact or respond, specifically:
  - > Users should not be able to post comments during children’s livestreams.
  - > Users should not be able to send gifts to children using the service through their livestreams.
  - > Users should not be able to post ’reactions’ during children’s livestreams.

> Viewers should be prevented from capturing or recording videos and images of children's livestreams via in-service functionalities. This should also be prevented via third-party services where technically feasible.

6.8 These functionality restrictions should be applied to all viewers of children in the UK's livestreams, regardless of whether or not they are logged into an account on the service.

6.9 To implement these restrictions in accordance with the measure we are proposing to include in the Illegal Content U2U Codes, services should use highly effective age assurance to apply these measures to users that have not been determined to be adults.

Effectiveness of the proposed measures at addressing risks

6.10 In chapter four, we set out evidence on the risks to children from both livestreaming and its supporting functionalities of posting comments and reactions, sending gifts and recording content, which amplify significantly the risk children face when livestreaming. We consider that our proposals would be effective as they would substantially reduce the ability of livestream viewers to interact with child livestreamers in the UK. This would reduce bad actors' ability to cause a range of types of harm to children, particularly but not limited to grooming. We therefore assess that the benefits of these measures are likely to be significant. They would remove significantly children's exposure to dynamic, high-pressure and real-time coercion, pressure to comply, and risk of immediate and severe harm in settings where they often currently have to rely on their own immediate judgement to avoid harm and keep safe.

6.11 Given the challenges in moderating livestreamed content as noted in paragraph 4.15 of chapter four, our assessment for consultation is that our proposed approach would afford children materially better protection than they currently have. We consider that this will be the case even once content moderation for livestreaming is made more robust through the implementation of Illegal Harms U2U Codes ICU C1 and ICU C2.[148] This is due to the scale of moderation required to protect children and the absence of proactive solutions to moderate content effectively in real time.

6.12 Because we propose that providers use highly effective age assurance, this should also ensure a robust approach to applying these settings to children's accounts.

6.13 In the following paragraphs we expand on how each element of our proposal will deliver benefits.

Removing comment functionality

6.14 Removing comment functionality would significantly reduce the risk to children from livestreaming by limiting the ability of bad actors to cause a range of harms including:

- grooming children who are livestreaming, including by encouraging them to move to more secure or encrypted services to further offend against them;

[148] ICU C1 recommends that service providers have a content moderation function to review and assess suspected illegal content. ICU C2 recommends that service providers have a content moderation function that allows for the swift take down of illegal content.

- coercing children who are livestreaming into sexual behaviour (which should also reduce the opportunities for perpetrators to record and further distribute such content, and so reduce children's vulnerability to further targeting and victimisation);
- encouraging suicide or self-harm of children who are livestreaming;
- posting hateful, bullying or denigrating comments alongside children's livestreams; and
- posting illegal content or content harmful to children alongside children's livestreams.

6.15 Removing comment functionality would also help reduce viewers' exposure to such content and reduce the negative influence on them through the risk of contagion and normalisation of harmful coping strategies (for example, in relation to suicide and self-harm as noted in paragraph 4.46 of chapter four).

Removing gifting functionality

6.16 The focus of this proposal is on preventing viewers of livestreams from sending gifts 'in the moment', which can be used to influence children's behaviour in real time. In-service gifting functionalities typically comprise virtual currencies or tokens created by service providers for users to send to each other within a specific service. They may also be able to be exchanged for an equivalent monetary value. We acknowledge that users may also seek to provide gifts, or that child livestreamers may invite gifts, outside of in-service gifting functionalities. Examples of this include links to external gifting sites, 'wishlists', or giving bank details. In-service gifting and use of virtual currencies and tokens are more frequently mentioned as being used to groom children and influence their behaviour 'in the moment' within our current evidence of harm as set out in chapter four. It is also likely to be more difficult for providers to prevent the sending of gifts to children outside of the service. Our proposals are therefore limited to restricting in-service gifting functionalities. However, we encourage service providers to consider how they can reduce the risks to children from third-party gifting. We also invite evidence on the extent to which third-party gifting is a risk to child livestreamers, and regarding what service providers can do to address it, to inform our future work.

6.17 We are proposing to define 'gift' in our Illegal Content U2U Codes as a 'user-to-user service functionality which enables a user to give a benefit (for example, money, in-service tokens, in-game or in-app gifts, or virtual currency) to another user'.

6.18 We assess that removing the ability to send gifts in-service while children livestream will reduce viewers' ability to entice or encourage children into sexual behaviour – and also potentially other forms of harm outlined in chapter four – in return for a reward.

Removing reaction features

6.19 'Reaction features' comprise the use of a user-to-user service functionality which allows users to express a reaction, such as approval or disapproval, of content that is shared by other users, through dedicated features that can be clicked or tapped by users. This for example includes 'liking' or 'disliking' a piece of content.[149]

[149] As set out in the glossary of: Ofcom, 2024. Illegal Harms Register of Risks. [accessed 12 June 2025]. This definition may include the following, as noted in section 4(1)(c) of the Online Safety Act 2023: (i) applying a

6.20 Removing reaction features would reduce the ability of bad actors to coerce, gamify or incentivise children to perform sexual acts, provide validation for thoughts of suicide or self-harm, and encourage children into other acts that are harmful to them.

Preventing content capture

6.21 As noted in chapter four, the majority of CSAM identified by the Internet Watch Foundation (IWF) derives from content captured from children's livestreams. Some U2U services have dedicated in-service content capture functionalities, and this should be applied for those functionalities. Users can often also capture content using third-party services, and this should also be prevented where it is technically feasible to do so.

6.22 We are proposing to define the term 'content capture' in our Illegal Content U2U Codes as follows: 'To record, copy or store content as it appears on the service. This can include, but is not limited to: the creation of clips from the content; downloading content; or photo, video or audio capture using device functionality or third-party software (for example, a screen shot/grab/recording)'.

6.23 While it is unlikely that content capture can be fully prevented (as, for example, perpetrators could record the footage using a separate device), the intention is that this measure will act as a friction to reduce opportunistic offending and help prevent the production of CSAM by making it harder to achieve. This should help reduce both the overall availability and onward sharing of CSAM generated via such means and repeat targeting by groomers, and so reduce the traumatisation and re-victimisation of children. This measure is also likely to reduce the sharing of recordings of livestreams in relation to other harms to children such as bullying.

6.24 The extent of the impact and change in behaviour will depend on the extent to which third-party content capture, in particular, can be prevented.

Risk of circumvention

6.25 There are various ways in which children and adults may be able to bypass our proposed measures. Examples include children using, or being encouraged by bad actors to use, forms of interaction outside of the livestreaming service such as messaging services, user profiles on other social media sites, and personal contact details. Children may seek to use the account of a person who has been determined to be an adult, if they have access to one, such as an account belonging to a family member. The risk to children is likely to increase where this occurs, particularly where such circumventions enable bad actors to identify children's real-world location or identity and contact them directly. However, we assess that overall, our proposed measures will reduce aggregate harms to children.

6.26 Bad actors are also likely to seek to contact children via direct messaging functionality within the service. We have not included direct messaging in our livestreaming proposals as we have already recommended that services at risk of grooming adopt ICU F1.4 of our Illegal Content U2U Codes. As noted in chapter 19, 'Increasing effectiveness for U2U settings, functionalities, and user support', we are also recommending the use of highly effective age assurance to apply this measure to children for the purpose of the Illegal

"like" or "dislike" button or other button of that nature, (ii) applying an emoji or symbol of any kind, (iii) engaging in yes/no voting, or (iv) rating or scoring the content (or the comments or reviews) in any way (including giving star or numerical ratings).

Content U2U Codes. ICU F1.4 sets out that if the service has direct messaging functionality, the provider should implement default settings ensuring that (1) if the service has user connection functionality, child user accounts can only receive direct messages from user accounts with which they have a specified connection, or (2) if the service does not have user connection functionality, users operating a child user account are provided with a means of actively confirming whether to receive a direct message sent from another user account before it is visible to them.[150]

6.27 We recommend that services take the approach set out in ICU F1.4, and also encourage them to consider if it is appropriate to allow direct message functionality to be available while children are livestreaming.

Technical feasibility

6.28 We assess that it is technically feasible for services to disable comments, gifting, reactions and content capture functionality on children's livestreams, where these are part of the service's own offering.

6.29 While preventing content capture in relation to third-party tools will almost certainly be more complex, there are several screen capture and recording prevention tools commonly used by video streaming, banking and some messaging services, such as flags offered via mobile operating systems. Our research suggests there are ways users may seek to bypass such measures, if they have sufficient technical knowledge. However, we consider that use of such technology would impede and create frictions for perpetrators' attempts to capture or record livestreamed content.

6.30 In recognition of these complexities, we are recommending that service providers only ensure that viewers of a child's livestream are unable to use tools outside of the service to screen capture or screen record where this is technically feasible. In contrast, we are recommending that in all circumstances providers should ensure that users viewing children's livestreams are unable to use the service to screen capture or screen record.

6.31 We welcome comments on the technical feasibility of these proposals in response to this consultation.

Other options considered

6.32 We considered and discounted a range of other options before deciding to propose our measures, including the following.

Disabling comments, reactions, and gifting as default for child users, but with the ability for older children to turn them back on

6.33 We considered this option to reflect the rights and evolving capacities of children as they get older, but discounted it for two primary reasons. Firstly, we do not currently have a

[150] ICU F1.4 states that this should apply unless direct messaging is a necessary and time critical element of another functionality, in which case before any interaction associated with that functionality begins, users operating a child user account should (1) be informed that they may receive direct messages from user accounts that are not connected to that child user account when using that functionality; and (2) having received that information, actively confirm that that they wish to proceed to use that functionality.

sufficient evidence base on the risk of harm for different age groups to make this recommendation for older children. Secondly, there are inconsistencies in asking services to take this approach when, so far, our Codes have aimed to establish a baseline level of protection for children of all ages. However, it is our view that there are increasingly effective means of using age assurance to differentiate between age groups. We also anticipate that improvements and new solutions will continue to develop at pace. We expect to continue gathering and analysing evidence on the capabilities of age assurance for determining the ages of children, in particular through our report on the use of age assurance, which we will publish in 2026.

Limiting the reach of child livestream broadcasts to friends lists or established connections on a service

6.34 We discounted this option as we assessed it could incentivise children to be less discerning about who they added as friends or connections to increase the reach of their livestreams. This approach could therefore also undermine the intention behind our Illegal Content U2U Codes grooming measures (ICU F1 and ICU F2).

Using highly effective age assurance to determine which users are adults and which are children and prevent them from interacting with each other on livestreaming services

6.35 We considered multiple different options to achieve this, but assessed all of them to be unworkable, primarily for the following reasons.

6.36 This approach would not afford child livestreamers protection from other children, who can represent a risk for CSEA, bullying, and other forms of harm. It could also give children a false sense of security, making them more vulnerable to abuse.

6.37 To prevent children and adults from interacting with each other, services would need to age gate the entire service and ensure all users complete an age check rather than just those who wish to broadcast a livestream, which could be considered disproportionate.

Impacts and Costs

Direct financial costs for service providers

6.38 For our proposed measures to only be implemented for child users, services will incur the cost of implementing highly effective age assurance if they do not already use it. Where a service has not already implemented highly effective age assurance, they will incur the full costs set out in paragraphs 19.26 to 19.40 of chapter 19 ‘Increasing effectiveness for U2U settings, functionalities, and user support’. In practice, some livestreaming services are likely to fall within scope of our age assurance measures under our Protection of Children U2U Codes.[151] These services should not incur additional costs to implement highly effective age assurance to comply with these measures.

6.39 There will be one-off implementation costs for all of our proposed measures, including designing required changes, engineering costs of testing and implementing those changes,

[151] PCU B4, PCU B5, PCU B6 and PCU B7.

and costs of further updates and refinement. Table 6.1 outlines the costs for each of the measures we are proposing.

Table 6.1: Summary of costs for measures to protect children on livestreaming services

Measure	Build cost	Ongoing maintenance cost
Disabling comment functions	£3,000 to £51,000	£1,000 to £13,000
Disabling gifting functionality	£10,000 to £41,000	£3,000 to £10,000
Disabling reaction-features	£8,000 to £31,000	£2,000 to £8,000
Preventing content capture	£8,000 to £31,000	£2,000 to £8,000
Total	£29,000 to £154,000	£8,000 to £39,000

Disabling comment functions

6.40 Existing measures ICU J2 and PCU J2 recommend that services provide their users with the option to disable comments on content they have posted.[152] We have used the same analysis for this measure, as services would need to implement the same comment disabling function, without the option for child users to turn it on. We assess that this would take one to ten weeks of software engineering time, with an equal amount of time from other professionals such as analysts, product managers and lawyers. This results in a one-off implementation cost of approximately £3,000 to £51,000, with annual maintenance costs of £1,000 to £13,000.[153] However, we expect costs to be lower for services which have already adopted either of those existing measures, as there is only the additional work needed for comments to be turned off permanently.

Disabling gifting functionality

6.41 We estimate that disabling in-service gifting functionalities would take four to eight weeks of software engineering time, with an equal amount of time from other professionals such as analysts, product managers and lawyers. This results in a one-off implementation cost of approximately £10,000 to £41,000 and annual maintenance costs of £3,000 to £10,000. If services also seek to prevent or reduce the likelihood of third-party gifting, then costs would likely increase significantly.

Disabling reaction features

6.42 We estimate that disabling reaction features on a service would take three to six weeks of software engineering time, with an equal amount of time from other professionals such as analysts, product managers and lawyers. Services could therefore incur implementation

[152] ICU J2 states that "the provider should offer every registered United Kingdom user the option of preventing any other users of the service from commenting on content posted on the service using their user account". PCU J2 recommends that children are provided the option to disable comments on their own posts.

[153] This uses our standard assumption that annual maintenance costs are equal to 25% (a quarter) of the initial build cost. We use this assumption for the other functionality restriction measures set out in this section. See Annex 15 for other details of our cost assumptions.

costs of approximately £8,000 to £31,000 and annual maintenance costs of £2,000 to £8,000.

Preventing content capture

6.43 We estimate that disabling in-service content capture functionalities would take three to six weeks of software engineering time, with an equal amount of time from other professionals such as analysts, product managers and lawyers. Services could therefore incur implementation costs of approximately £8,000 to £31,000 and annual maintenance costs of £2,000 to £8,000. If services were to look to also limit third-party content capture, we anticipate that there would be significant additional costs, including those associated with procuring external screen capture and recording prevention tools and labour costs of implementing them on a service.

Indirect effects on service providers

6.44 Overall, we consider that there may be impacts on the market and service providers from our measures, particularly where these may act as a barrier to growth or disincentive to entry for new services. This may entrench the market positions of larger services through the impact on service providers' revenues or business models where they rely on engagement with livestreaming and developing a user base. Given the severity of the harm we are seeking to mitigate and the risk of real-time coercion, we consider these wider potential impacts to be proportionate. Services will be able to limit the application of these measures to a limited group of users for whom the potential harm is experienced (child users).

6.45 We consider that the evidence supports that this is a minimum standard of safety required to keep child users safe when broadcasting livestreams and therefore consider potential wider market impacts detailed further below justified. Evidence suggests that livestreaming services tend to be medium or larger in size, as the fixed costs of establishing a livestreaming service are likely to be significant. The application of these measures to all livestreaming services limits the potential for distortionary impacts on the market. Finally, the industry may adapt in response to these measures to find new ways to attract and engage users and the overall market impact is uncertain.

6.46 There is a risk that some service providers may exit the market as a result of our proposed measures or remove livestreaming functionality for UK users. This may be more likely for services where livestreaming is a secondary or lower priority part of their service provision, and may be more likely for services that do not use highly effective age assurance for any other part of the service. Alternatively, rather than adopt our measures, some service providers may choose to prevent children from livestreaming entirely as a means of managing the risk.

6.47 The number of child broadcasters on livestreaming services may reduce significantly as livestreaming becomes less popular once interaction functionalities are removed. The number of viewers of children's livestreams may also reduce due to the lack of ability to interact with those undertaking the livestream. Both of these would reduce services' revenue. Livestreaming-only services, especially those with a large proportion of younger users, would be particularly impacted. Our proposed measures may therefore impact on livestreaming services' business model and the revenue that such services receive through advertising.

6.48 Where a service has a recommender system based on user interactions (such as ‘likes’), this may have an impact on how children’s livestreams are treated within the service – for example, by affecting where they appear in rankings or whether they are suggested to other viewers. Some services may therefore need to adapt their recommender systems.

Negative impacts on users

6.49 The most adverse impact of our proposals may be on older children, such as 16- to 17-year-olds, who may feel unduly limited in their ability to receive interaction from and respond to others via various functionalities. This impact would also be felt by adult users who choose not to undergo an age assurance process that service providers will have in place on their services.

6.50 Children (and adults who do not go through highly effective age assurance to be determined as adults) may experience reduced financial gain from livestreaming through our proposed limitations on in-service gifting. These represent only one form of revenue for livestreamers, as income may also be derived from other sources, including subscriptions, viewer numbers, advertisements and sponsorships.[154] It is possible these may also decline due to reduced visible engagement with their content. Given the potential for harm outlined in chapter four, our provisional assessment is that this approach is a necessary and proportionate means of addressing the risks to children. While it may not appeal to everyone affected, users may be able to offset any reduced income through other forms of online or offline activity. The potential loss of income will also increase adults’ willingness to undergo an age check, reducing the negative impacts on them from receiving a limited livestreaming experience.

6.51 While limiting interaction functionalities should help prevent bad actors, it will also likely remove opportunities for children to receive positive feedback from well-meaning viewers and reduce engagement with and viewing of livestreams, including by their friends and trusted contacts. As noted in chapter four, through this consultation and in parallel with it, we are working to obtain further insights from children and those that care for them, experts, and those with lived experience of online harms on the benefits (and risks) of livestreaming. A primary aim of doing so is to increase our understanding and inform our thinking to minimise the negative impacts of our proposed measures on children.

6.52 There may be negative impacts on legitimate purposes for content capture, as referred to within the ‘Freedom of expression and association’ section, paragraph 6.64.

Rights assessment

6.53 We consider that our proposed measures for inclusion in the Illegal Content and Protection of Children U2U Codes would have the following impacts on rights.

[154] The available information on financial gain from livestreaming is limited and varies by source. However, anecdotal evidence suggests that while some livestreamers generate a large amount of revenue, most do not. For example, analysis of a leak of Twitch data in 2021 suggested that 85% (over 8 in 10) of streamers earned less than $10,000 over two years: ronin.achikochi, 2021. Diving into the Twitch leaked revenue dataset. [accessed 13 June 2025]. Further anecdotal evidence provides a breakdown of earnings on Twitch based on numbers of viewers: Streamerfacts, 2024. How much do Twitch streamers make in 2025? Small vs Big vs Top-tier. [accessed 13 June 2025].

Freedom of expression and association

6.54 In chapter three we set out the detail of the rights to freedom of expression and association, including that we must exercise our functions under the Act in a way that does not restrict these rights unless we are satisfied that it is necessary and proportionate to do so.

6.55 These measures will make it more difficult to perpetrate grooming offences and capture CSAM, and to generate and disseminate other kinds of priority illegal content. This will reduce child users' vulnerability to the same. Any interference should be weighed against this.

6.56 There may be some positive impacts on children's freedom of expression and association rights if these measures are able to ensure safer online spaces in which children are confident about joining and expressing themselves with reduced risks to their safety.

6.57 We also acknowledge that the use of highly effective age assurance in the implementation of these measures may impact users' rights of expression and association.

Disabling comments, gifting, and reaction functions

6.58 Preventing children from receiving interactions from viewers of their livestreams by disabling comments, gifting and reaction functionalities, and preventing viewers from sending the same presents a significant restriction of users' freedom of expression and association rights.

6.59 We acknowledge that this limits child livestreamers' potential opportunities to build connections and community in online spaces, to receive well-meaning interactions and to monetise their content in some cases. This may have a particular impact on older children, for example 16- to 17-year-olds, who may feel unduly limited in their ability to receive interaction from and respond to others via the restricted functionalities. However, child livestreamers remain able to livestream and users are able to view their livestreams. Some services also enable livestreamers to post their livestreams as videos after the livestream has ended; in these cases it may also be possible for other users to interact with these videos.[155] It also remains possible for users to be part of online communities of shared interests in other ways.

6.60 The evidence set out in chapter four indicates that these functionalities enable a wide range of harms, including grooming offences and CSAM. Therefore their removal will reduce the harm caused.

6.61 Other measures and guidance provide safeguards, including our Part 3 Highly Effective Age Assurance (HEAA) Guidance, and measure ICU B1 (which defines highly effective age assurance and recommends that services consider certain matters when implementing highly effective age assurance). Also of relevance are measures ICU D15 and ICU D16 which concern age assessment appeals which are complaints about the incorrect assessment of a user's age.

6.62 Based on our analysis as detailed in this section, we consider that the interference with rights is likely to be proportionate.

[155] Subject to other relevant measures, for example ICU J2 'Disabling comments'.

Preventing content capture

6.63 We consider preventing content capture of child users' livestreams presents a limited restriction to users' freedom of expression rights as it does not impact their ability to view or broadcast livestreams. Depending on the functionality offered by the service, child livestreamers may be able to record or capture their own livestreamed content (for example, for onward sharing once the livestream has concluded).

6.64 We acknowledge that this may present a restriction of the freedom of expression rights of users who wish to capture parts of livestreams for legitimate purposes, such as journalism. However, this will not occur in all circumstances; rather it will only occur when the person creating the livestream has not been determined to be an adult through the use of highly effective age assurance. Given the strong evidence that capturing livestreams can be a way to generate new kinds of priority illegal content, particularly CSAM (see paragraph 4.67 to 4.68 of chapter four), we consider this restriction of rights to be proportionate. Where a livestreamer is determined to be an adult through the use of highly effective age assurance, other users will still be able to record/capture their livestream (where this is enabled by the service or where third-party record/capture tools are used).

6.65 Recording or capturing livestreams without the broadcaster's consent may, in some cases, already be contrary to providers' terms of service. In such cases, the additional interference with viewers' rights is minimal.

6.66 Where it is not technically feasible for services to prevent capture using third-party tools, these tools will still be available to users who would like to capture livestreams in the exercise of their own freedom of expression (provided this is permitted by services' terms).

6.67 Any interference with viewers' freedom of expression is to be balanced against the substantial public interest in addressing illegal content. In particular, it should be balanced against the substantial public interest in addressing CSAM, and the way in which the measure contributes to protecting the privacy and other rights of victims depicted in CSAM derived from livestreams.

Privacy and data protection

6.68 Article 8 of the European Convention on Human Rights (ECHR) sets out the right to respect an individual's private and family life. The use of an age assurance process to determine if a user is a child will involve the collection and processing of personal data.

6.69 All methods of age assurance will involve the processing of some personal data of individuals. We are recommending that highly effective age assurance is implemented. This will have an impact on users' right to privacy and their rights under data protection law. However, we consider that the impact on users' privacy rights is likely to be proportionate to the benefits of using highly effective age assurance. This is because for our measures to be effective at reducing harm to children it is important that providers are able to correctly and reliably apply the measures to children using the service. Our view is that using highly effective age assurance is likely to secure this objective. Further, we note that service providers are required to comply with relevant data protection legislation when processing personal data; this helps to safeguard against disproportionate interferences with user privacy.

6.70 The degree of interference will depend on several factors, including the nature of the information required to complete the age assurance process. The more sensitive the information required, the more intrusive the method of age assurance is likely to be.

6.71 We considered recommending that providers apply the restrictions described in this chapter to all users (and not recommending the use of highly effective age assurance). We recognise that this would have achieved our objective of protecting children while avoiding the interference with user privacy associated with highly effective age assurance. However, this would have impacted the freedom of expression rights of a greater number of users. In other words, applying the restrictions to all users has a greater impact on freedom of expression but a lesser impact on privacy, while using highly effective age assurance to apply the measures at children has a greater impact on privacy but a lesser impact on freedom of expression. In light of the severity of potential harm, we consider that the use of highly effective age assurance for implementing these measures is both appropriate and proportionate.

6.72 We expect providers to comply with relevant data protection legislation when processing personal data for the purposes of implementing highly effective age assurance. We also expect providers to consider relevant Information Commissioner's Office (ICO) guidance, and the ICO Children's Code. Providers should also consider the Part 3 HEAA Guidance when implementing highly effective age assurance to assist in protecting users' privacy. The Part 3 HEAA Guidance includes relevant information about the data protection regime and relevant ICO guidance. Section 5 of the Part 3 HEAA Guidance[156] sets out further details about privacy and data protection in relation to highly effective age assurance.

6.73 Other measures and guidance provide safeguards against disproportionate interferences with privacy. In particular, measure ICU B1 recommends that providers should ensure that their age assurance process meets the criteria of technical accuracy, robustness, reliability and fairness to ensure that it is highly effective. These measures also recommend that providers take certain matters into account when implementing HEAA, including that it is easy to use, should work effectively for users regardless of their characteristics or whether they are members of a certain group, and the age-appropriate design code and the Information Commissioner's opinion regarding the same. These measures are supported by our HEAA Guidance. Lastly, measures ICU D15 and D16, which concern complaints about the incorrect assessment of a user's age.

6.74 This interference is balanced against substantial public interest in addressing CSEA and other harms to children, and the way in which the measures contribute to protecting the privacy and other rights of victims depicted in CSAM derived from livestreams, and other users who have, for example, had their livestreams captured without their knowledge.

Which providers should implement this measure

6.75 Given the level of risk related to livestreaming, we propose that the measure included in the Illegal Content U2U Code should be applied to all user-to-user services offering one-to-many livestreaming, where it is possible for children to access the service, or a part of it. This includes services that only provide livestreaming and those where it is a secondary feature.

[156] Ofcom, 2025. Guidance on highly effective age assurance for Part 3 Services [accessed 16 June 2025]

6.76 We have proposed to apply the measures to services with a livestreaming functionality, rather than to services with a particular risk-level, type of harm or by service size because:

- a range of illegal content and content harmful to children can impact both children and other users when livestreaming, beyond child sexual abuse and grooming;
- many of these harms are enabled through livestreaming functionality specifically, and other functionalities typically built into it such as comments, reactions and gifting;
- there are likely to be very few if any services with a livestreaming functionality not at heightened risk of relevant harms; and
- if any services are not currently at heightened risk, we assess the risk of users migrating to them to be high if other services are required to implement these measures, meaning the risk on those services would likely therefore increase.

Provisional conclusion

6.77 On balance, we consider that, given the severity of the risks and our current evidence base, our proposed measures would be a proportionate intervention which would make a material contribution to reducing the risk of harm to children from one-to-many livestreaming.

6.78 As reflected in our consultation questions, we request that respondents submit evidence on the effectiveness of our proposals and potential alternative measures.

6.79 We will consider stakeholder responses and evidence. Our provisional view is that the measures we have proposed are the minimum intervention necessary to address the harms we have identified. However, if respondents identify alternative approaches which are both proportionate and equally or more effective than the proposals set out in this chapter, we are willing to consider them.

7. Livestreaming – further steps

7.1 We believe that the proposed measures set out in chapters 5 and 6 will form a baseline of protection for users from the risks associated with livestreaming. However, we also consider that further action may be needed to fully address the risks that livestreaming poses, both to adults and to children.

7.2 Therefore, during the consultation process, we will initiate a programme of stakeholder engagement to ascertain whether there are other options that may further help protect users, including children, from the risks of livestreaming.

Options for going further – all users

7.3 We are aware that several service providers already deploy preventive safety by design methods to protect users to make it more difficult to broadcast illegal content via livestream.

7.4 Examples of these methods include:

- Age of account: Users are unable to access parts of the service during a defined period of time after registering the account
- Number of connections: Users are unable to access parts of the service until they reach a minimum number of connections, friends or followers.
- Community Guidelines: Users are unable to access parts or all of the service for a period of time if they have breached the service's community guidelines or terms of service.
- Identity verification: Users can only access parts of the service once they have completed an identity verification process which may involve them providing relevant information such as their name and address.

7.5 As part of the stakeholder engagement described above, we plan to explore the advantages and disadvantages of these methods, so we can decide whether there is a case for adding them to our Codes.

7.6 Further, we understand that some providers already signpost users at risk of suicide to crisis prevention information. For example, Meta states that when it becomes aware of users posting about suicide and self-harm, it provides them with resources and links to local organisations that could offer support.157 Samaritans recommend that when providers remove suicide or self-harm content, they should explain to the user where to find support.158 We also understand that some providers notify emergency services when a user is at risk of suicide or harming themselves. For example, Meta state that if it identifies that someone is at immediate risk of harming themselves, it contacts emergency services.159

7.7 We currently lack evidence regarding how these measures are implemented by service providers and its effectiveness specifically in relation to livestreaming. At this stage we are

[157] Meta, 2025 Suicide Prevention. [accessed 4 June 2025].
[158] Samaritans, 2020. Managing self-harm and suicide content online. [accessed 14 May 2025].
[159] Meta, 2025. Suicide, self-injury and eating disorders. [accessed 3 June 2025]

therefore not in a position to propose further measures in our Codes. We do however encourage providers in scope of this measure to signpost to crisis prevention information and to notify emergency services when livestreaming users are at risk of taking their life or harming themselves. We will keep this position under review and may revisit our position as our evidence base expands including through responses to this consultation.

Options for going further – child users

7.8 We will review the responses to this consultation, as well as any other relevant information, including further insights we obtain from children and those that care for them, experts, and those with lived experience of online harms, on the benefits and risks to children from livestreaming.

7.9 If we receive compelling evidence during the consultation, we are prepared to go further, which could include recommending that children are prevented from livestreaming entirely. We note that this is a step that some services have already taken as a matter of their own service design.160

7.10 Equally, if respondents identify alternative approaches which are both proportionate and demonstrably equally or more effective than the proposals set out in chapter 6, we are willing to consider them.

7.11 If we proceed with our proposed measures, we will monitor how effective they are on an ongoing basis. We will also review the efficacy of alternative measures service providers are using to protect children.

[160] For example, in 2019, Live.me raised the age at which users could broadcast livestreams on the service to 18 (Source: Melugin, B., 2019. Live streaming app 'LiveMe' makes major changes following award-winning FOX 11 investigation. FOX 11 Los Angeles, 3 October. [accessed 4 June 2025]). TikTok did the same in 2022 (Source: TikTok, 2022. Enhancing the LIVE community experience with new features, updates, and policies [accessed 4 June 2025]).

8. Proactive Technology – Introduction

What is Proactive Technology?

8.1 We use the term ‘proactive technology’ to describe a set of technologies that can be used as part of a moderation system. The technologies (which are defined in the Online Safety Act) are: content identification technology, user profiling technology, and behaviour identification technology (we explain these terms further below).

8.2 In practice, this includes rules-based software such as hash matching or keyword detection, Automated content classifiers (algorithmic systems which analyse content and associated signals), and other proactive technologies which analyse inputs – including network analysis, user profiling and behavioural analysis.

Why does it matter?

8.3 The regulatory framework set up by the Online Safety Act is designed to make life online safer for UK users by reducing the risks posed by illegal content and content harmful to children. Proactive technology is an important tool that service providers can (and in many cases already do) use to assist in identifying this content. Without using proactive technology, providers are reliant on methods like user reports, or human moderation. Given the volumes of content on some large services, we consider that these methods are unlikely to be sufficient to tackle the volume of potential illegal and/or content harmful to children on a service.

8.4 On this basis, the Act allows Ofcom to make recommendations about the use of proactive technology as part of our Codes of Practice, subject to certain constraints discussed below. Our existing Codes already recommend the use of proactive technology to tackle CSEA, via perceptual hash matching for known image-based CSAM and direct URL detection for CSAM URLs. We are now proposing additional measures which would recommend the use of proactive technology.

The Online Safety Act and Proactive Technology

Types of proactive technology

8.5 As detailed above, the Act defines proactive technology as content identification technology, user profiling technology or behaviour identification technology. These are defined as follows:

- Content identification technology: technology which analyses content to assess whether it is content of a particular kind. Examples of this technology include algorithms, keyword matching, image matching, or image classification (for example, illegal content).[161]

- User profiling technology: technology which analyses (any or all of) relevant content, user data, or metadata relating to relevant content or user data for the purposes of building a profile of a user to assess characteristics such as age.[162]
- Behaviour identification technology: technology which analyses (any or all of) relevant content, user data, or metadata relating to relevant content or user data to assess a user's online behaviour or patterns of online behaviour (for example, to assess whether a user may be involved in, or be the victim of, illegal activity).[163]

Constraints on Ofcom's power to recommend proactive technology

8.6 Because of the speed at which proactive technology can assess large volumes of content, it is important it is used responsibly and in a way which doesn't impinge on the legitimate rights of users. Paragraph 13 of Schedule 4 of the Act contains several constraints on our ability to recommend the use of proactive technology in Codes of Practice.

8.7 Ofcom may not recommend in a Code of Practice the use of proactive technology to analyse user-generated content communicated "privately", or metadata relating to that content.[164] Accordingly, each of the measures in chapters 9-13 applies only to content communicated publicly by means of the service. However, it is open to service providers to decide to use proactive technology in relation to content communicated privately by means of the service, including to detect illegal content, and there may be good reasons for them to choose to do so in some cases.

8.8 The Act requires Ofcom to have regard to the degree of accuracy, effectiveness and lack of bias achieved by a technology in deciding whether to recommend it in a Code of Practice or use it as part of a confirmation decision.[165] This helps minimise disproportionate impacts on privacy and freedom of expression. The Act also empowers Ofcom to set out principles in a Code of Practice designed to ensure that proactive technology or its use is (as far as possible) accurate, effective and free of bias.[166] Our assessment of these factors is explained in the following chapters, 9-13.

8.9 A proactive technology measure may be applied to services of a particular kind or size only if Ofcom is satisfied that the use of the technology in question by such services would be proportionate to the risk of harm that the measure is designed to safeguard against (taking

[161] Content identification technology is not regarded as proactive technology if it is used in response to a report from a user or other person about particular content.

[162] Technology which analyses data specifically provided by a user for the purposes of the provider verifying or estimating the user's age in order to decide whether to allow the user to access a service (or part of a service) or particular content is not regarded as user profiling technology.

[163] This technology is not regarded as proactive technology if it is used in response to concerns identified by another person or an automated tool about a particular user.

[164] Schedule 4 of the Act, paragraphs 13(4). We have set out guidance which is intended to assist services in deciding whether content is communicated 'publicly' or 'privately' for this purpose here: Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act.

[165] Schedule 4 to the Act, paragraph 13(6).

[166] Schedule 4 to the Act, paragraph 13(6).

into account, in particular, the risk profile relating to such services).[167] Our assessment of proportionality is set out in each of the chapters to follow.

8.10 The Act imposes other requirements on services in connection with their use of proactive technologies. For example, a service provider's duty to include provisions in their terms of service giving information about any proactive technology used for the purpose of compliance with the safety duties.

What the following chapters cover

8.11 Chapters 9-13 set out details of five new proposed measures that would make use of proactive technology.

8.12 The first two measures relate to what we are calling a 'principles-based' proactive technology measure. They recommend that providers of in-scope services assess whether proactive technology that is sufficiently accurate, effective and free from bias exists that could be used to identify a range of harms on their service (CSEA, fraud, illegal suicide content, and primary priority content that is harmful to children), and if so, to deploy that technology; and to assess whether existing proactive technology used on their service is sufficiently accurate, effective and free from bias.

8.13 We have also produced draft guidance to support providers to conduct assessments of proactive technology; this is included as Annex 6.

8.14 In addition, we are suggesting an amendment to the illegal content judgements guidance (ICJG) for CSAM. This is to guide providers on how to make judgements on content in circumstances where it may be technically feasible to review a name, icon or bio / description of a message, group chat or forum, but where they cannot review the content of the message, group chat or forum. This is specific to the context of perpetrators using names, icons or bios / descriptions to indicate to other perpetrators that CSAM content can be found within.

8.15 There are also three measures which recommend the use of a specific type of proactive technology called 'hash matching'. Hash matching refers to a process for detecting content which has previously been identified as being illegal or otherwise violative to prevent the upload and sharing of that content. It involves creating a hash (an identifying series of characters) of a unique piece of content, storing that hash in a database, and using an algorithm to detect attempts to upload the same or similar versions of that content. Once a match of some form is established, the content can either undergo review - which may include human review - or be removed automatically.

8.16 There are two basic forms of hash matching:

- **Perceptual hash matching** determines whether a given file is similar to a hashed file by creating a hash of the given file and comparing it to hashes of known reference files. A threshold is set to determine when there is sufficient similarity between the hashes for the files to be considered a (near) match. This allows modifications from an original file to be detected.
- **Cryptographic hash matching** only detects files which are an exact match. Perpetrators may slightly alter images to circumvent cryptographic hash matching. Perceptual hash

[167] Schedule 4 to the Act, paragraph 13(5).

matching mitigates this risk as it picks up a wider range of potential near matches, making circumvention more difficult, and can therefore be more effective in detecting illegal content. However, there is an increased probability that legal content may be incorrectly detected as a match for illegal content (described as a 'false positive').

8.17 In our Illegal Content User-to-user Codes, we recommended that certain user-to-user service providers use perceptual hash-matching to detect image-based child sexual abuse material (CSAM).

8.18 This type of proactive technology can be used to detect a range of priority illegal content. As such, we are proposing to also introduce hash-matching measures for:

a) Intimate Image Abuse content; and
b) Terrorism content.

8.19 In addition, after careful consideration, we are also proposing a minor change to an existing Codes measure which recommends hash-matching technology for CSAM, to increase the number of high-risk service providers that are in scope of the measure.

8.20 Details of these proposals are set out in Chapters 9-13.

9. Proactive Technology

Summary

Technology that can scan and assess content (which we call ‘proactive technology’) provides an important method for keeping people safe online. Service providers use it to detect content at scale that is suspected to be illegal content, or content harmful to children, as part of their wider content moderation system. At the same time, proactive technology is better at detecting some types of content than others. For example, it is materially better at identifying content such as child sexual abuse material (CSAM) than at identifying content where a detailed understanding of context is necessary to make judgments.

We are proposing that service providers should assess whether accurate and effective proactive technology exists for detecting content including CSAM, grooming, fraud and financial services offences, suicide, self-harm and eating disorder content and pornography. Where such tools exist, we propose service providers should use them as part of their content moderation systems for publicly communicated content and metadata. [168] We also propose a measure for service who already use these systems.

These proposals should increase the amount of illegal content and content harmful to children which service providers detect and action accordingly. At the same time, our proposals are designed to safeguard freedom of expression because they only recommend that service providers use proactive technology when it is accurate, effective and free from bias.

We propose these measures should apply to large user-to-user services that are medium risk and user-to-user services with more than 700,000 monthly UK users that are high risk of at least one of the following relevant harms[169]:

- Illegal harms: image-based CSAM, CSAM URLs, grooming, fraud and other financial services offences, encouraging or assisting suicide; and
- Content harmful to children: primary priority content, which includes content relating to suicide, self-harm, eating disorders and pornography.[170]

We also propose that these measures should apply to:

- File storage and file sharing services of any size that are high risk of image-based CSAM to align with the measure already in our Illegal Harms User-to-user Codes on CSAM hash matching.
- All user-to-user services with a high risk of grooming.

[168] This measure applies to: a) providers who are not already deploying proactive technology to detect or support the detection of all the target illegal content and/or content harmful to children they are at risk of; and b) providers who have already deployed proactive technology to detect or support the detection of all the target illegal content and/or content harmful to children they are at risk of, and have concluded that this proactive technology cannot meet the proactive technology criteria within a reasonable timeframe.

[169] We refer to the harms listed in this paragraph and as set out in ICU 11.2, ICU 12.2, PCU 9.1 and PCU 10.1 as 'relevant harms' for the purpose of these measures.

[170] Only services likely to be accessed by children and are large user-to-user services that are medium-risk, and user-to-user services with more than 700K users that are high-risk should consider Content Harmful to Children.

Our proposals

Numbers	Proposed measure	Who should implement this
ICU C11	Assessing proactive technology for use to detect or support the detection of target illegal content.	• Large user-to-user services that are medium or high risk for at least one relevant harm • User-to-user services with more than 700,000 monthly UK users that are high risk, for at least one relevant harm • User-to-user services that are file-storage and file-sharing services which identify a high risk of image-based CSAM, regardless of size • All user-to-user services which identify a high risk of grooming
PCU C9[171]	Assessing proactive technology for use to detect or support the detection of target content harmful to children.	
ICU C12	Assessing existing proactive technology for use to detect or support the detection of target illegal content.	
PCU C10[172]	Assessing existing proactive technology for use to detect or support the detection of target content harmful to children.	

Consultation questions

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

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

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

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

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

Structure of the chapter

9.1 This chapter is structured as follows:

- Definitions
- What is proactive technology?
- What risk does the use of proactive technology seek to address
- Our proposals
- Interaction with the regulatory framework
- Measure ICU C11 and PCU C9
- Measure ICU C12 and PCU C10

[171] This measure is limited to services likely to be accessed by children.

[172] This measure is limited to services likely to be accessed by children.

- Other options considered

9.2 The sections which describe Measure ICU C11 and PCU C9 and ICU C12 and PCU C10 are structured as follows:

- Explanation of the measure
- Effectiveness at addressing risks and benefits associated with the measure
- Impact and costs
- Rights assessment
- Who should implement this measure
- Provisional conclusion

9.3 This chapter relates to the following Annex:

- Annex 13: Further evidence of relevant harms

In the chapter, we use the following terms:

We use the term 'proactive technology' as defined in the Online Safety Act 2023 (the Act).[173] Broadly speaking, 'proactive technology' may refer to content identification technology, user profiling technology, or behaviour identification technology. We define each of these in greater detail, by reference to the Act, as follows:

- **Content identification technology:** technology, such as algorithms, keyword matching, image matching, or image classification, which analyses content to assess whether it is content of a particular kind (for example, illegal content).[174]
- **User profiling technology:** technology which analyses (any or all of) relevant content, user data, or metadata relating to relevant content or user data for the purposes of building a profile of a user to assess characteristics such age.[175]
- **Behaviour identification technology:** technology which analyses (any or all of) relevant content, user data, or metadata relating to relevant content or user data to assess a user's online behaviour or patterns of online behaviour (for example, to assess whether a user may be involved in, or be the victim of, illegal activity).[176]

We refer to the range of **inputs** that proactive technology may analyse (such as relevant content and, in the case of user profiling and behaviour identification technology, user data and metadata relating to relevant content or user data) as 'inputs'.

Proactive technology system: a content moderation system that includes proactive technology. It can operate with partial or no human involvement. For the avoidance of doubt, this term refers to the overall system, including any human review element.

[173] Section 231 of the Act.

[174] Content identification technology is not to be regarded as proactive technology if it is used in response to a report from a user or other person about particular content.

[175] Technology which analyses data specifically provided by a user for the purposes of the provider verifying or estimating the user's age in order to decide whether to allow the user to access a service (or part of a service) or particular content and does not analyse any other data or content is not regarded as user profiling technology.

[176] This technology is not regarded as proactive technology if it is used in response to concerns identified by another person or an automated tool about a particular user.

Relevant content: This includes:

- **regulated user-generated content**[177] that may be encountered by UK users (including child users) of the service by means of the service and is communicated publicly by means of the service;[178] or
- any material which, if it were present on the service, would be such content.

Target illegal content means **relevant content** that:

a) amounts to an offence in relation to the relevant harms; or
b) is illegal content proxy,[179] where the provider is satisfied that its terms of service prohibit **the relevant harms**.

Target content harmful to children means **relevant content** that either:

a) relates to one of the following:
  (i) where the service is a large service, the specific kinds of primary priority content for which the service is at medium or high risk, to the extent they are relevant primary priority content; or
  (ii) where the service is not a large service, the specific kinds of primary priority content for which the service is at high risk, to the extent they are relevant primary priority content; or
b) is **content that is harmful to children proxy**[180], where the provider is satisfied that its terms of service prohibit the specific kinds of relevant primary priority content.

Target illegal content and/or content harmful to children: target illegal content and/or target content harmful to children.

What is proactive technology?

9.4 The term 'proactive technology' is defined in the Act.[181]. Service providers commonly use proactive technology for a range of purposes, including keeping users safer by identifying illegal content and/or content harmful to children. See chapter 8, 'Proactive Technology - Introduction' for further information.

9.5 As explained in chapter 8 'Proactive Technology - Introduction: The Online Safety Act and Proactive Technology', Ofcom can recommend that services use proactive technology as

[177] Section 55 of the Act.

[178] Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act [accessed 13 June 2025].

[179] Illegal content proxy: In the Illegal Content User-to-user Codes, we defined "illegal content proxy" as content that a provider determines to be in breach of its terms of service, where: the provider has reason to suspect that the content may be illegal content; and the provider is satisfied that its terms of service prohibit the type of illegal content which it had reason to suspect existed.

[180] Content that is harmful to children proxy (referred to in this chapter as 'content harmful to children proxy'): In the Protection of Children User-to-user Code, we define "content that is harmful to children proxy" as primary priority content (PPC) proxy, priority content (PC) proxy or non-designated content (NDC) proxy". This is content that a provider determines to be in breach of its terms of service, where: a) the provider had reason to suspect that the content may be relevant PPC, PC and/or NDC; and b) the provider is satisfied that its terms of service prohibit the type of relevant PPC, PC and/or NDC which it had reason to suspect existed.

[181] Section 231 of the Act.

part of our Codes of Practice but must follow the rules set out in the Act regarding proactive technology.

9.6 The Act requires Ofcom to have regard to the degree of accuracy, effectiveness and lack of bias achieved by a technology in deciding whether to include it as a proactive technology measure in a Code of Practice or in a confirmation decision.[182] This helps minimise disproportionate impacts on privacy and freedom of expression (for example, instances of content being wrongly removed due to proactive technology).

9.7 While the concepts of accuracy, effectiveness and lack of bias are inherently linked, we have broadly defined these for the purpose of these proposed measures as follows:

- **Accuracy:** the technical correctness and reliability of the proactive technology as assessed during development, testing, and evaluation, including the extent to which errors are minimised through evidence-based methodologies and appropriate performance metrics.
- **Effectiveness:** the utility and ability of the proactive technology to achieve intended outcomes, evaluated by assessing its real-world impact, alignment with specific goals, and operational relevance for the intended use case.
- **Lack of bias:** The extent to which the proactive technology avoids discrimination and negative impacts on equity and fairness, including with respect to the treatment of different users and the handling of different types of content or other inputs.

9.8 The Act also empowers Ofcom to set out principles in a Code of Practice designed to ensure that proactive technology or its use is (as far as possible) accurate, effective and free from bias.[183]

9.9 In this chapter, we set out a principles-based approach when recommending proactive technology to detect or support the detection of target illegal content and/or content harmful to children.

9.10 In having regard to accuracy, effectiveness and lack of bias, we have developed proactive technology criteria that we propose should be met when providers are sourcing, developing, and/or assessing existing proactive technology. These criteria are designed to ensure that proactive technology is sufficiently accurate, effective and free from bias while giving providers the flexibility to assess and, if appropriate, deploy the right proactive technology for their service.

9.11 We have proposed that providers use proactive technology to detect or support the detection of specific harms because they align closely with our eight targets for immediate action,[184] and we have a higher degree of confidence that proactive technology that is accurate, effective and free from bias is likely to be available for these harms. We provide further detail about how we selected these harms in the section below ‘Other options considered’.

[182] Schedule 4 to the Act, paragraph 13(6).

[183] Schedule 4 to the Act, paragraph 13(6).

[184] For our eight targets for immediate action see Ofcom, 2024, Implementing the Online Safety Act: progress update, p.8 [accessed 13 June 2025.]

9.12 We developed the proactive technology criteria drawing on evidence from relevant national and international frameworks on responsible artificial intelligence (AI) and automated content moderation.[185]

9.13 In devising the proactive technology criteria, we have considered common challenges in the design and deployment of detection technologies (such as accuracy, fairness, oversight, and adaptability) and translated these into a practical, outcomes-focused set of criteria.

9.14 While the proactive technology criteria were not adopted from a single external standard, they reflect widely recognised concerns and are consistent with guidance and recommendations published by academic, regulatory, and industry bodies.[186] [187] [188] [189] We consider this process ensures the criteria, when taken as a whole, are proportionate, technology-agnostic, and adaptable to a range of service types and harms.

9.15 The proactive technology criteria are as follows:

Table 9.1: Proactive technology criteria and description

Criteria	Description
Use of high-quality data	The proactive technology has been developed and tested using high-quality[190] datasets appropriate to and reflecting a broad range of inputs relevant to the harm[191] it is intended to detect (as identified in the service's risk assessment(s)).
Addressing biases	Potential biases have been identified and addressed during the design and development process, and risks are appropriately managed and addressed throughout the proactive technology's lifecycle.
Evaluating performance	The proactive technology has been evaluated using appropriate performance metrics and configured so that its performance strikes an appropriate balance between precision[192] and recall.[193] [194]

[185] See footnotes 20-24.

[186] NIST, 2023. NIST, 2023, NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence | NIST [accessed June 13 2025.]

[187] DSIT, 2024, A pro-innovation approach to AI regulation [accessed 13 June 2025.]

[188] OECD, 2023, AI principles | OECD [accessed 13 June 2025.]

[189] EU 2024, Regulation - EU - 2024/1689 - EN - EUR-Lex [accessed 13 June 2025.]

[190] "High quality" refers to data that is accurate, legally and ethically sourced, well-labelled (where appropriate), representative of the harm and context the proactive technology is intended to address and sufficiently diverse to support meaningful evaluation, including (where relevant) testing the proactive technology's ability to detect content it has not previously encountered. The degree to which a dataset may be considered "high-quality" will vary based on the harm being addressed and the context in which the proactive technology is used. However, indicators that a dataset might not be considered "high quality" would include (for example) unclear sourcing, evidence of poor or inconsistent labelling, or limited representation of relevant harms or input types.

[191] Or a relevant proxy, if applicable.

[192] Precision is the proportion of identified cases that are true positives.

[193] Recall is the proportion of true positive cases that are correctly identified.

[194] Please see section 'Additional Factors to consider for 'Evaluating Performance' and 'Human Review' for certain factors providers should consider related to the testing, configuration, deployment and ongoing monitoring of the proactive technology.

Criteria	Description
Safeguards against misuse and exploitation	Safeguards are in place to identify and appropriately manage security threats and risks of exploitation and misuse, including through the use of access restrictions[195] and system integrity protections.
Contextual testing and evaluation	The proactive technology's performance has been evaluated in real-world use cases relevant to the provider's content (having regard to the risk of harm to individuals identified in the providers' risk assessment(s)) and the results indicate it correctly detects the harm it is intended to detect. This includes testing for scalability, handling of different media types (where relevant), and whether the proactive technology's performance could be improved by layering with complementary approaches or (in the case of existing deployments) by updating to a more current version.
Maintenance and ongoing monitoring	Mechanisms are in place to monitor and maintain the proactive technology's effectiveness over time, including processes for regular review and iterative adjustments to respond to emerging circumvention techniques, biases, or new content types.
Human review	Policies and processes are in place for human review and action is taken in accordance with that policy, including the evaluation of outputs during development (where applicable), and the human review of an appropriate proportion of the outputs of the proactive technology during deployment. Outputs should be explainable to the extent necessary to support meaningful human judgement and accountability.[196]
Incorporating feedback	Feedback mechanisms are in place to maintain or improve performance over time. This includes updating the proactive technology with diverse and up-to-date datasets to reflect evolving trends or emerging types of illegal content and/or content harmful to children and/or integrating ongoing feedback from users and individuals working in content moderation[197] into its development, while managing the risk of introducing additional bias.

9.16 Where the proactive technology meets the proactive technology criteria, we consider that it is sufficiently accurate, effective and lacking in bias for us to recommend its deployment.

[195] Access restrictions" refers to limitations placed on interaction with system operations or data, based on the roles or responsibilities assigned to any entity (including individuals, devices, or systems such as software components, applications, and automated processes), whether internal and external. This aligns with principles on controlling access, as outlined in the UK Cyber Assessment Framework, and supported by international standards such as ISO/IEC27001 and NIST SP-800-53.

[196] Please see section 'Additional Factors to consider for 'Evaluating Performance' and 'Human Review' for certain factors providers should consider related to the testing, configuration, deployment and ongoing monitoring of the proactive technology.

[197] For an explanation of 'individuals working in content moderation' see our December 2024 Illegal Harms Statement: Volume 2: Service design and user choice pp. 79-80 [accessed 12 June 2025].

Additional factors to consider for 'Evaluating performance' and 'Human review'

9.17 For the criteria 'Evaluating performance' and 'Human review', providers should consider certain factors related to the testing, configuration, deployment and ongoing monitoring of the proactive technology.

9.18 For criteria '**Evaluating performance',** when configuring the technology so that it strikes an appropriate balance between precision and recall, providers should ensure that the following matters are taken into account:

- the service's risk of relevant harm(s) proposed as part of this measure, reflecting the risk assessment(s) of the services and any information reasonably available to the provider about the prevalence of target illegal content and/or content harmful to children on the service.
- the proportion of detected content that is a false positive.
- the effectiveness of the systems and/or processes used to identify false positives; and
- in connection with CSAM or grooming, the importance of minimising the reporting of false positives to the National Crime Agency (NCA) or a foreign agency.

9.19 For criteria '**Human review'**, when determining what is an appropriate proportion of detected content to review by humans, providers have flexibility to decide what proportion of detected content it is appropriate to review, however in so doing, providers should ensure that the following matters are taken into account:

- the principle that the resource dedicated to review of detected content should be proportionate to the degree of accuracy achieved by the technology and any associated systems and processes.
- the principle that content with a higher likelihood of being a false positive should be prioritised for review; and
- in the case of CSAM or grooming, the importance of minimising the reporting of false positives to the NCA or a foreign agency.

9.20 Further guidance on the proactive technology criteria, including examples of how to assess proactive technology against the criteria and how they can be met, are set out in the 'Proactive Technology Draft Guidance'.

What risk does the use of proactive technology seek to address?

9.21 The proposed measures capture the following harms:

- illegal harms: image-based CSAM, CSAM URLs, [198] [199] grooming, fraud and other financial services offences (fraud), encouraging or assisting suicide (suicide); and
- content harmful to children: primary priority content (PPC) which includes pornographic, suicide, self-harm and eating disorder content.

9.22 We note that there are various types of CSAM in-scope of the proposed measures:

- For services with the applicable risk level for image-based CSAM, we recommend they deploy proactive technology to detect or support the detection of this, including unknown image-based CSAM. This is image-based CSAM that has not previously been identified (and is not captured by our hash matching measure, ICU C9). [200]
- For services with the applicable risk level for CSAM URLs, we recommend they deploy proactive technology to detect or support the detection of this, including disguised/altered CSAM URLs. These are CSAM URLs that have been obfuscated to evade detection (and are not captured by our direct URL detection measure, ICU C10).

9.23 In either of the above cases, we recommend services deploy proactive technology to detect 'CSAM discussion'. CSAM discussion refers to CSAM in the form of written material or messages (other than CSAM URLs).[201] While services do not assess their risk of CSAM discussion, we are recommending that they deploy technology to detect it or support its detection if they are in-scope of the recommendations to detect, or support the detection of, image-based CSAM and/or CSAM URLs. This is because text-based discussion often occurs alongside image-based CSAM and CSAM URLs and will aid in its detection.

9.24 Our evidence suggests the harms captured by these measures are prevalent across a broad range of services and have a significant adverse impact on UK users.[202] For more details on the relevant offences and how service providers can assess whether content amounts to illegal content, see the Illegal Content Judgements Guidance; for details on content harmful to children, see the Guidance on Content Harmful to Children.[203]

9.25 The impact of these harms, including psychological, physical and financial impacts, are widespread and can be devastating. Our research, along with other evidence set out in the Illegal Harms Register of Risks (Illegal Harms Register) and the Children's Register of Risks

[198] These measures apply to unknown image-based CSAM (image-based CSAM that has not previously been identified) and disguised/altered URLs (harms not captured by ICU9 and ICU10). Please see 'What risk does the use of proactive technology seek to address?' for further detail.

[199] We also refer to 'CSAM discussion' in the measure as a type of CSAM that providers should detect, which is associated with a service's risk level for image-based CSAM and CSAM URLs. Please 'What risk does the use of proactive technology seek to address?' for further detail.

[200] Illegal content Codes of Practice for user-to-user services, pp.24-31 [accessed 13 June 2025].

[201] In our codes, CSAM is defined as content that amounts to certain offences specified in Schedule 6 to the Act. Some of these offences relate to content which may be in text form, for example, offences of encouraging and assisting the commission of CSAM offences, conspiring to carry out CSAM offences, or material which is an obscene article encouraging the commission of other child sexual exploitation and abuse offences. The term 'CSAM discussion' refers to such text-based CSAM content.

[202] Ofcom, 2024, Ofcom Illegal Harms Register of Risks, Section 2b Child Sexual Abuse Material (CSAM) [accessed 13 June 2025.].

[203] Ofcom, 2024, Illegal Content Judgements Guidance (ICJG) [accessed 13 June 2025.]; Ofcom, 2025, Guidance on content harmful to children [accessed 13 June 2025.].

(Children's Register), shows how such harms may manifest online and the scale and impact of these harms.[204]

9.26 In our November 2023 Consultation on Protecting People from Illegal Harms Online (November 2023 Consultation) we received responses from some stakeholders who were concerned about providers' reliance on user reporting to identify illegal content and supported proactive content detection.[205] For example, one stakeholder suggested pre-screening content to prevent CSAM.[206] Another stakeholder also highlighted the need for measures to include proactively detecting certain forms of priority illegal content, such as suicide and self-harm content.[207]

9.27 In our May 2024 Consultation on Protecting Children from Harms Online (May 2024 Consultation) we acknowledged that although we considered recommending specific automated technology, our view was that specific proactive technologies by themselves were not sufficiently sophisticated to accurately detect content harmful to children. However, we said that we planned an additional consultation on how automated detection tools can be used to mitigate the risk of content harmful to children.[208]

9.28 In response to the May 2024 Consultation, however, we received responses from many stakeholders calling for the Codes to recommend automated content moderation measures.[209] This included suggestions that we should recommend proactive detection of content harmful to children.[210] We are now acting on this feedback.

Interaction with the regulatory framework

Illegal content safety duties

9.29 Part 3 of the Act places duties on providers of regulated services. These include duties set out in Section 10 of the Act that require providers of regulated user-to-user services to take

[204] Ofcom, 2025, Ofcom Children's Register of Risks. [accessed 13 June 2025].

[205] Institute for Strategic Dialogue response to November 2023 Consultation, p.10; Marie Collins Foundation response to November 2023 Consultation, pp.9-10; Online Safety Act Network (OSA Network) Annex B response to November 2023 Consultation, pp.13-14. We note that the Commissioner Designate for Victims of Crime Northern Ireland made a similar point in its response to the May 2024 Consultation, p.5.

[206] Marie Collins Foundation response to November 2023 Consultation, pp.9-10.

[207] Molly Rose Foundation response to November 2023 Consultation, p.35.

[208] Ofcom, 2024, Consultation: Protecting Children from Harms Online p.16 [accessed 13 June 2025].

[209] Canadian Centre for Child Protection (C3P) response to May 2024 Consultation, pp.18-19, 21; Children and Young People's Commissioner Scotland response to May 2024 Consultation, p.8; Christian Action Research and Education (CARE) response to May 2024 Consultation, p.5; Conscious Advertising Network and Middleton K., University of Portsmouth response to May 2024 Consultation, pp.8-9; Nexus response to May 2024 Consultation, p.15; NSPCC response to May 2024 Consultation, pp.48, 62; Office of the Children's Commissioner for England response to May 2024 Consultation, p.47; OSA Network (1) response to May 2024 Consultation, p.73; Parenting Focus response to May 2024 Consultation, pp.36-37; and, UK Safer Internet centre (UKSIC) response to May 2024 Consultation, p.34.

[210] Amaran, M. response to May 2024 Consultation, p.3; Centre to End All Sexual Exploitation (CEASE) response to May 2024 Consultation, p.17; Commissioner Designate for Victims of Crime Northern Ireland response to May 2024 Consultation, p.5; Conscious Advertising Network and Middleton K., University of Portsmouth response to May 2024 Consultation, p.27; Google response to May 2024 Consultation, p.23; NCA response to May 2024 Consultation, p.9; and, Vodafone response to May 2024 Consultation, p.2.

or use proportionate measures relating to the design or operation of the service to, among other things:[211]

- prevent individuals from encountering priority illegal content;
- effectively mitigate and manage the risk of the service being used for the commission or facilitation of a priority offence, as identified in the most recent illegal content risk assessment of the service;
- effectively mitigate and manage the risks of harm to individuals, as identified in a service's most recent illegal content risk assessment;
- minimise the length of time for which any priority illegal content is present; and
- swiftly take down illegal content when the provider becomes aware of it.

9.30 Priority illegal content includes content that amounts to an offence specified in Schedule 6 (CSEA offences) or Schedule 7.

Children's safety duties

9.31 The Act also places duties on providers of services likely to be accessed by children to (1) take steps to prevent and protect children from encountering content harmful to children, including through content moderation where proportionate, and (2) take proportionate measures to effectively mitigate and manage the risk and impact of harm to children from content that is harmful to children.[212]

9.32 PPC that is harmful to children includes pornographic content;[213] content which encourages, promotes or provides instruction for suicide; content which encourages, promotes or provides instruction for an act of deliberate self-injury; and content which encourages, promotes or provides instruction for an eating disorder or behaviours associated with an eating disorder.

Proactive technology

9.33 We have defined 'proactive technology' in chapter 8.

9.34 We propose two measures addressing the relevant harms. Subject to the outcome of this consultation, both measures will be included in the Illegal Content User-to-user Codes and the Protection of Children User-to-user Code,[214] respectively.

9.35 Whether services are in scope of one or both measures will be determined by the size and risk of the service.[215]

[211] Section 10(2)-(3) of the Act

[212] Section 12 of the Act

[213] Pornographic content is content other than content within s61(6) of the Act.

[214] The measures included in the Protection of Children User-to-user Code will be limited to services likely to be accessed by children.

[215] This is determined by its illegal content risk assessment, the children's access assessment and the children's risk assessment (where relevant). See Ofcom, 2024, Risk Assessment Guidance and Risk Profiles; Ofcom, 2025, Children's Access Assessments Guidance; and, Ofcom, 2025, Children's Risk Assessment Guidance and Children's Risk Profiles [accessed 13 June 2025].

Our proposals

9.36 Given the high volume of content on user-to-user services, we do not consider that relying solely on complaints and human moderation would be sufficient to detect the relevant harms at the scale needed to ensure users are protected from content that is suspected to be target illegal content and/or content harmful to children.

9.37 Therefore, we consider that proactive technology is essential for detecting these types of content at scale. The use of such technology can support swift and accurate intervention to protect UK users.

9.38 In the Illegal Content User-to-user Codes and the Protection of Children User-to-user Code, we recommended that providers review, assess, and swiftly take action on illegal content, illegal content proxy, content harmful to children and/or content harmful to children proxy when they become aware of it and where it is currently technically feasible to do so (Measures ICU C1 and C2 and Measures PCU C1 and C2). [216]

9.39 In the Illegal Content User-to-user Codes, we also recommended that providers should ensure that hash-matching technology is used to detect and remove CSAM (ICU C9) and that providers should detect and remove content communicated publicly on the service which matches a URL on a list of URLs previously identified as hosting CSAM (ICU C10).

9.40 However, we did not recommend specific measures on the use of proactive technologies for a wider range of illegal content and content harmful to children. Having considered evidence from the November 2023 Consultation and the May 2024 Consultation, we are now proposing the following two additional measures for inclusion in the Illegal Content User-to-user Codes:

Number	Proposed measure	Who should implement this
ICU C11	Assessing proactive technology for use to detect or support the detection of target illegal content. .	• Large user-to-user services that are medium/high risk for at least one relevant harm; • User-to-user services with more than 700,000 monthly UK users that are high risk for at least one relevant harm; • User-to-user services that are file-storage and file-sharing services which identify a high risk of image-based CSAM, regardless of size. • All user-to-user services which identify a high risk of grooming.
ICU C12	Assessing proactive technology for use to detect or support the detection of target content harmful to children.	

Corresponding measures for inclusion in the Protection of Children User-to-user Code:

Number	Proposed measure	Who should implement this

[216] Illegal content Codes of Practice for user-to-user services, p.19-20 and Protection of Children Code of Practice for user-to-user services, pp.29-34. [accessed 13 June 2025].

PCU C9	Assessing existing proactive technology for use to detect or support the detection of target illegal content.	• Large user-to-user services, likely to be accessed by children, that are medium/high risk for at least one relevant harm; • User-to-user services with more than 700,000 monthly UK users, likely to be accessed by children, that are high risk for at least one relevant harm;
PCU C10	Assessing existing proactive technology for use to detect or support the detection of target content harmful to children.	

9.41 To ensure that the measure is proportionate, we are only proposing to recommend the measure for the following harms:

- illegal harms: image-based CSAM, CSAM URLs, grooming, fraud and other financial services offences (fraud), encouraging or assisting suicide (suicide); and
- content harmful to children: primary priority content (PPC) which includes pornographic, suicide, self-harm and eating disorder content.

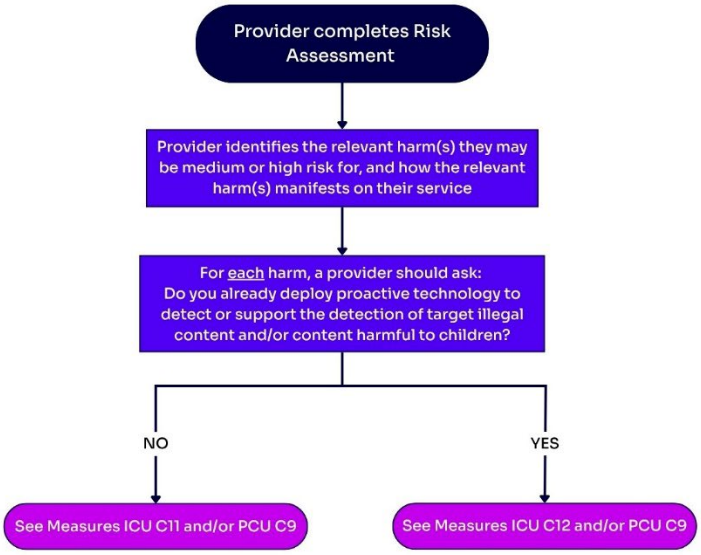
9.42 We are not recommending these measures for all priority illegal content, priority content (PC) and non-designated content (NDC) at this time please see 'Proactive technology for all priority illegal harms' and 'Proactive technology for PC and NDC' in the 'Other options considered' section of this chapter for further explanation. However, where providers currently deploy proactive technologies for a wider range of harms, we encourage them to continue to do so and to consider the proactive technology criteria where appropriate.

9.43 We recognise that the harms in scope of these measures are broad and may manifest in a variety of ways across different services. However, after completing their risk assessment(s), we expect providers to understand the harm(s) they may be medium or high risk for and how the harm(s) manifests on their service. Providers should take this into account when assessing whether they can deploy proactive technology that meets the proactive technology criteria.

9.44 We consider that these measures will deliver significant benefits to UK users. We expect they will increase the volume of target illegal content and/or content harmful to children detected on their service and the speed at which it is detected.[217]

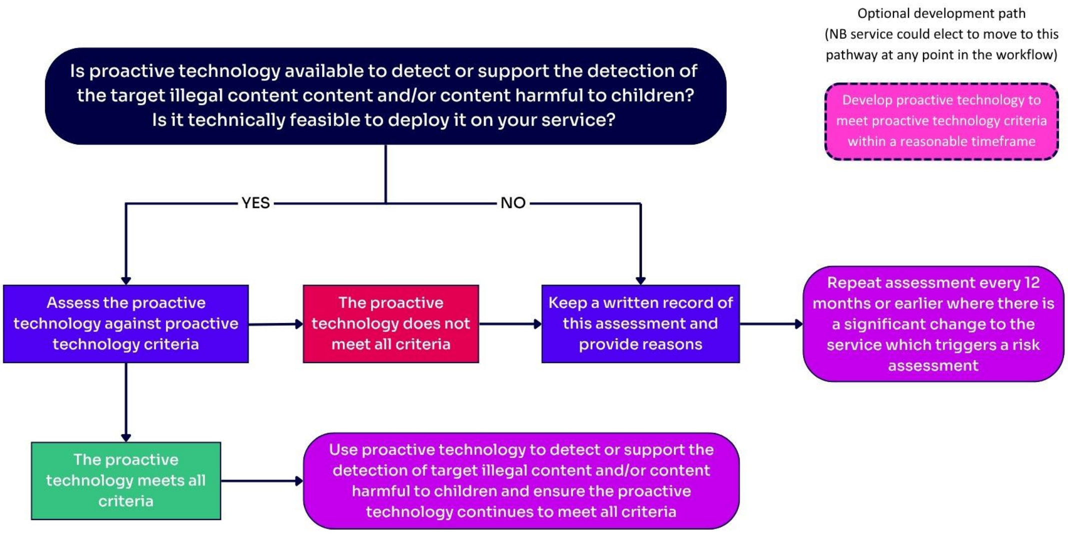
[217] In our Protection of Children Code of Practice for user-to-user services, PCU E1 recommends that providers exclude content indicated potentially to be primary priority content from recommender feeds. In this consultation, we are also proposing to introduce a similar measure into our Illegal content User-to-user Codes (see chapter 14: Recommender systems). In both cases, service providers should use 'relevant available information' to make this assessment. We define 'relevant available information' to include 'indicators generated by technology used on the service'. Where service providers detect content using proactive technology in line with our proposals in this chapter, this is an 'indicator generated by technology used on the service' and should therefore be treated as 'relevant available information' for the purpose of our recommender systems measures.

Diagram 9.1: Understanding which measures apply to service providers


Provider completes Risk Assessment
Provider identifies the relevant harm(s) they may be medium or high risk for, and how the relevant harm(s) manifests on their service
For each harm, a provider should ask: Do you already deploy proactive technology to detect or support the detection of target illegal content and/or content harmful to children?
NO
YES
See Measures ICU C11 and/or PCU C9
See Measures ICU C12 and/or PCU C9

Measure ICU C11 and PCU C9: assessing proactive technology for use to detect or support the detection of target illegal content and/or content harmful to children

Diagram 9.2: Flowchart visualising Measures ICU C11 and PCU C9


Optional development path
(NB service could elect to move to this pathway at any point in the workflow)
Is proactive technology available to detect or support the detection of the target illegal content content and/or content harmful to children? Is it technically feasible to deploy it on your service?
Develop proactive technology to meet proactive technology criteria within a reasonable timeframe
YES
NO
Assess the proactive technology against proactive technology criteria
The proactive technology does not meet all criteria
Keep a written record of this assessment and provide reasons
Repeat assessment every 12 months or earlier where there is a significant change to the service which triggers a risk assessment
The proactive technology meets all criteria
Use proactive technology to detect or support the detection of target illegal content and/or content harmful to children and ensure the proactive technology continues to meet all criteria

Explanation of the measure

Proposed framework for compliance

9.45 We propose that service providers should:

- Assess whether proactive technology to detect or support the detection of target illegal content and/or content harmful to children is available, is technically feasible to deploy on their service, and meets the proactive technology criteria.
- Deploy proactive technology to detect or support the detection of target illegal content and/or content harmful to children so that it continues to meet the proactive technology criteria.

9.46 We expect some providers may source proactive technology from third-party providers. For details on how providers should source proactive technology, see ‘Sourcing proactive technology’.

9.47 Some providers may decide to develop proactive technology for the relevant harms in-house. For more details, see ‘Developing proactive technology’.

9.48 Where providers currently deploy proactive technologies for the relevant harm, see ‘Measure ICU C12 and PCU C10: Assessing existing proactive technology for use to detect or support the detection of target illegal content and/or content harmful to children’ for actions we recommend they should take.

Sourcing proactive technology

Step 1: Assess whether proactive technology to detect or support the detection of target illegal content and/or content harmful to children is available, is technically feasible to deploy on their service, and meets the proactive technology criteria.

9.49 In this section we expand on each of the elements of the assessment process in turn.

Identify whether proactive technology that detects or supports the detection of target illegal content and/or content harmful to children is available

9.50 The harms captured in this measure may manifest in a variety of ways across different services.

9.51 Providers should identify whether proactive technology that detects relevant harms, in the way they manifest on their service, is available.

9.52 When considering the relevant harms, providers may consider identifying whether proactive technology exists in relation to an illegal content proxy or content harmful to children proxy. For example, a service that has a high-risk of pornographic content may prohibit nudity in their terms of service. In such cases, the service may consider whether proactive technology exists to detect nudity specifically, rather than pornographic content. We consider this approach would align with how providers decide to operate their content moderation function.[218]

9.53 Different types and/or combinations of proactive technology can be used to detect or support the detection of target illegal content and/or content harmful to children, using different techniques such as content analysis, network analysis, metadata analysis, user profiling, and behaviour analysis. When making their assessment, providers should consider what type(s) of proactive technology would be most appropriate to address the relevant harms on their service.

9.54 For the harms captured by this measure, we have a higher degree of confidence that proactive technology that is accurate, effective and free from bias is likely to be available. However, where providers conclude that no proactive technology that detects the relevant harms is available, they should keep a record of this decision explaining how the decision was made. For further information, see ‘Where providers cannot deploy proactive technology’.

Identify whether it is technically feasible to implement that proactive technology on their service

9.55 The proposed measure will only apply where it is technically feasible for a service to implement proactive technology. We do not consider that it would be technically infeasible to implement proactive technology merely because to do so would require some changes to be made to the design and/or operation of the service.

[218] See measures: ICU C1, Having a content moderation function to review and assess suspected illegal content; ICU C2, Having a content moderation function that allows for the swift take down of illegal content; PCU C1, Having a content moderation function to review and assess suspected content that is harmful to children; and PCU C2, Having a content moderation function that allows for swift action against content harmful to children.

9.56 However, the proposed measure will not apply to providers for whom it is not technically feasible to analyse user-generated content present or disseminated on the service to assess whether it is content of a particular kind, particularly where such changes as would need to be made to enable this would materially compromise the security of the service.

9.57 Where providers conclude that it is not technically feasible to implement the proactive technology on their service, they should keep a record of this decision together with reasons why. For details, see 'Where providers cannot deploy proactive technology'.

9.58 Where a service provider claims that it is technically infeasible for it to implement proactive technology, we may investigate this. Should we then find that it is technically feasible for the provider to implement proactive technology, the measure will apply to them. We do not consider that technical limitations will necessarily remain on an ongoing basis.

9.59 Any provider that can currently deploy proactive technology to analyse inputs on the service and seeks to amend its technical architecture to make it infeasible for it to do so will trigger the statutory requirement for a new risk assessment, before any change is made, in line with its duties under the Act and Ofcom's risk assessment guidance. The provider will need to transparently explain to Ofcom the nature of the risks arising from this decision and plans for mitigating these risks.

Assess whether proactive technology meets the proactive technology criteria

9.60 We do not propose to be prescriptive as to the specific type or combination of proactive technology that should be deployed.[219] We understand that the way harm manifests on services will vary. Therefore, providers should deploy proactive technology that addresses the relevant harms in a way that is appropriate for their service, while ensuring that the proactive technology is accurate, effective, and lacking in bias.

9.61 Providers should assess whether available proactive technology can be deployed in a way that meets the proactive technology criteria.

9.62 The proactive technology criteria include criteria that relate to both the proactive technology itself, and the way it is configured and deployed by the service provider.

9.63 Where proactive technology meets all the criteria, we consider that the proactive technology is sufficiently accurate, effective, and lacking in bias; and therefore, it is proportionate to recommend that the provider deploy it in accordance with the other elements of this measure.

9.64 Further guidance on the proactive technology criteria, including how providers may assess technology against them, is set out in the 'Proactive Technology Draft Guidance'.

9.65 Where a provider concludes that the proactive technology does not meet the proactive technology criteria, they should keep a record of this decision and provide their reasons. Where a provider claims that the proactive technology does not meet the proactive

[219] Different types of proactive technology may be deployed together to produce a more accurate and effective overall proactive technology system. For example, Google uses automated technology alongside other systems and processes to detect content. Google's response to our formal information request. February 2025. Proactive technology that identifies behavioural signals in addition to content identification technology may contribute to better identification of target content uploaded by users perceived to be more high-risk.

technology criteria, we may investigate this. For details, see 'Where providers cannot deploy proactive technology'.

Step 2: Deploy proactive technology to detect or support the detection of target illegal content and/or content harmful to children so that it continues to meet the proactive technology criteria

9.66 Where proactive technology is available, is technically feasible to implement on the service, and meets the proactive technology criteria, the provider should procure and deploy it.

9.67 We propose that providers should implement proactive technology to detect or support the detection of target illegal content and/or content harmful to children before or as soon as is practicable after it can be encountered by UK users, having regard to the desirability of minimising the number of UK users encountering target illegal content and/or content harmful to children.

9.68 We consider that providers should scan inputs that are generated on, uploaded to or shared on the service (or that a user seeks to generate, upload, or share) after the technology is implemented. To the extent that this involves scanning content, this should take place before or as soon as practicable after that content can be encountered by UK users of the service.[220]

9.69 When deploying proactive technology to detect or support the detection of target content harmful to children, we consider that providers should ensure that relevant inputs that are generated on, uploaded to or shared on a child-accessible part of the service (or that a user seeks to so generate, upload or share) after the technology is implemented are analysed before or as soon as practicable after target content harmful to children can be encountered by relevant users.

9.70 We are recommending a different approach for services which are in scope of this measure because they are at the applicable risk level for image-based CSAM, CSAM URLs or grooming. As noted above, these services should deploy proactive technology to detect or support the detection of:

a) image-based CSAM (including unknown image-based CSAM) and CSAM discussion (in the case of services at the applicable risk level for image-based CSAM).
b) CSAM URLs (including altered/disguised URLs) and CSAM discussion (in the case of services at the applicable risk level for CSAM URLs); and
c) Grooming

9.71 For these services, we recommend that providers should:

- scan inputs that are generated on, uploaded to or shared on the service (or that a user seeks to generate, upload, or share) after the technology is implemented; and
- also scan pre-existing content (content present on the service before the proactive technology is implemented).

[220] For proactive technology that analyses metadata and user/behavioural signals, we understand that, for the technology to be as effective as possible, it may need to be trained from data that existed before the technology was implemented (for example, training from historic user reports or user networks). We do not discourage providers from doing this or using third-party technology that operates in this way; however, we do not expect the proactive technology itself to be applied to pre-existing content once it has been trained/implemented.

9.72 We know that CSAM may circulate for years after it is first uploaded.[221] The repeated circulation of CSAM is a source of continued traumatisation and re-victimisation, and this is not lessened by how old the content is. Regarding grooming, children can take many years to report these cases, if they do so at all, leading to significant underreporting of these offences online.[222] For these reasons, we propose that services at risk of image-based CSAM, CSAM URLs, and/or grooming should scan all content, including pre-existing content, on their service.

9.73 Aside from CSAM and grooming, where providers consider it appropriate, they may deploy proactive technology to scan pre-existing content as well as inputs that are generated on, uploaded to, or shared on the service after the technology is implemented. We encourage them to do so, for example, when detecting fraudulent content.[223] However we are not recommending this within the measure.

9.74 Where proactive technology detects or supports the detection of illegal content and/or content harmful to children, providers should treat this as reason to suspect that the content may be target illegal content and/or content harmful to children. Providers should therefore take appropriate action in line with existing content moderation measures, namely ICU C1 and ICU C2 (in the Illegal Content User-to-user Codes of Practice) and PCU C1 and PCU C2 (in the Protection of Children User-to-user Code of Practice), as applicable.

Where providers cannot deploy proactive technology

9.75 Where providers cannot deploy proactive technology that meets the proactive technology criteria for the relevant harm(s), they should keep a record of this decision.[224]

9.76 The record, which should be kept in line with Ofcom's record-keeping and review guidance, should include:[225]

- the steps the provider took to reach the decision, including the methodology used, the evidence considered, and the results of the assessments undertaken.

[221] For example, the National Centre for Missing and Exploited Children (NCMEC) state in their 2024 report that CSAM of one child "has been circulated for the past 19 years appearing more than 1.3 million times in submissions" to the organisation. NCMEC, 2024, NCMEC 2024 CyberTipline Report [accessed 06 May 2025].

[222] For example, results of an online survey with male child sexual abuse survivors showed that the length of time until first disclosure ranged from 0 to 63 years with a mean of more than two decades (21.45 years). Easton, S., 2019. Childhood disclosure of sexual abuse and mental health outcomes in adulthood: Assessing merits of early disclosure and discussion. Child Abuse and Neglect, 93, 208-214.

[223] In the case of fraudulent content, scanning pre-existing content could improve the capability of proactive technologies to contextualise 'new' content, thereby increasing the accuracy and effectiveness of the technology. This may be the case in instances where pre-existing content may be used to enable fraud, for example sharing of 'how-to guides' on how to commit fraud of financial services offences, or other articles for use in fraud.

[224] Providers may also take alternative measures to comply with safety duties. 'Alternative measures' means measures other than measures which are (in relation to the provider and the service in question) applicable measures in a Code of Practice. If a provider decides to take alternative measures, they have a duty to keep certain records. The details of these records are set out in our Ofcom, 2025, Record Keeping and Review Guidance, and in Annex 16 [accessed 13 June 2025].

[225] Ofcom, 2024. Record Keeping and Review Guidance – the written record should be durable, accessible, easy to understand and up to date [accessed 13 June 2025].

- justification for the decision (for example, an explanation as to why the provider considers it is not technically feasible to deploy proactive technology on the service, or the specific criteria that could not be met and why).
- any other actions the provider is undertaking to proactively identify these harms on their service; and
- the governance and sign-off process.

9.77 Providers should continue to consider whether proactive technology can be deployed on their service to detect relevant harm(s) as conditions change. For example, where new technology becomes available, new information suggests that a proactive technology that did not previously meet the criteria may do so now, or there is evidence of new and increasing relevant harm(s) on their service.

9.78 We propose providers should conduct this assessment at a minimum of every 12 months, or earlier where there is a change to any aspect of the service's design or operation which would amount to a significant change and therefore trigger a risk assessment under section 9(4) of the Act.[226] A record of the results of these reviews should be kept as outlined in paragraph 9.76.

Developing proactive technology

9.79 Some providers, particularly larger providers, currently deploy proactive technology that they have developed in-house. Some providers choose a hybrid approach, relying on a combination of third-party and in-house solutions.

9.80 Providers may develop proactive technology and still be compliant with this measure provided that it meets all the proactive technology criteria.

9.81 When making this decision providers may wish to initially consider the factors set out in the Proactive Technology Draft Guidance, see 'Proactive Technology Draft Guidance'.

9.82 If a provider decides to develop in-house proactive technology that meets the proactive technology criteria, they should do so within a reasonable time. In such cases, it is not necessary for the provider to carry out step 1 of section 'Sourcing proactive technology' (assess whether proactive technology is available, is technically feasible to deploy on their service, and meets the proactive technology criteria), however, they should deploy the proactive technology so that it continues to meet the proactive technology criteria (see step 2).

9.83 Alternatively, if, at any stage, a provider is unable to and/or decides not to develop proactive technology that meets the proactive technology criteria, they should return to step 1 in section 'Sourcing Proactive Technology'.

Benefits and effectiveness at addressing risks

9.84 Given the severity of the relevant harms and the high volume of content that is generated, uploaded and shared on some user-to-user services, we consider proactive technology to be an essential way of detecting or supporting the detection of target illegal content and/or content harmful to children once services reach a certain size. At this scale, proactive

[226] See, Ofcom, 2024, Risk Assessment Guidance and Risk Profiles and Ofcom, 2025, Children's Risk Assessment Guidance and Children's Risk Profiles for when risk assessment(s) should be reviewed and updated [accessed 13 June 2025].

technology plays a crucial role in detecting and taking action on content in a way that is sufficiently quick and accurate to prevent widespread harm to UK users.

9.85 While user reporting is important, we acknowledge that it has some limitations as a means of detecting illegal content and/or content harmful to children for the following reasons.

- First, users are not always able to identify such harms, for example, evidence suggests that children generally do not recognise grooming.[227]
- Second, although users may identify illegal content and/or content harmful to children, evidence suggests that often, they will not report this type of content.[228] [229]
- Finally, in some instances, illegal content and/or content harmful to children will circulate in particular communities for example, offender communities, such as perpetrators of CSAM, or suicide forums. These groups do not have incentives to report this type of content.

9.86 Once deployed, proactive technology should enable quicker detection of more content than would be possible if relying on human reviewers or reporting mechanisms alone.[230] [231] Consequently, we consider that the measure would increase the amount of illegal content which services detect and remove, and content harmful to children which services take action on accordingly, thereby reducing users' exposure to such content and delivering significant benefits to UK users.

9.87 Evidence suggests that the largest and most technologically developed providers currently make use of a variety of proactive technologies to detect or support the detection of target illegal content and/or content harmful to children.

9.88 Analysis of the Digital Services Act (DSA) Transparency Database demonstrates the important role played by automated technologies in moderating content at scale. For example, from the August 2023 to October 2023 reporting period, approximately 90 million (61%) content moderation decisions were fully automated by the largest providers.[232]

9.89 There are several third-party providers that supply proactive technology to detect the harms captured by this measure. For example, Thorn's automated classifier 'Safer' is used by providers to detect suspected unknown CSAM. Thorn reported that over 1.5 million of the 3.8 million CSAM files detected in 2023 were predicted as new or previously

[227] Quayle, E., Jonsson, L., Lööf, L., 2012. Online behaviour related to child sexual abuse. Interviews with affected young people. [accessed 22 May 2025]; Council of the Baltic Sea States, Stockholm: ROBERT project. [accessed 22 May 2025]; Katz, C., Piller, S., Glucklich, T., & Matty, D. E., 2021. "Stop Waking the Dead": Internet Child Sexual Abuse and Perspectives on Its Disclosure. Journal of Interpersonal Violence, 36 (9–10), NP5084–NP5104 [accessed 22 May 2025].

[228] Ofcom, 2024. Research: Joint online, calls and texts fraud research January 2024, p.22 Nationwide, 2024. Press Statement: "Not worth it": why scams under £100 go unreported. [accessed 22 May 2025].

[229] In a survey by Project deSHAME, 39% of children reported that they would 'ignore' online harassment. They also reported a number of barriers to seeking help, such as being 'too embarrassed' (52%) or being 'worried about what would happen next' (42%). Project deSHAME, 2017. Young people's experiences of online harassment. Available at: Research report | Childnet [accessed 28 May 2025].

[230] Ferro, C., Gsenger, R., Kübler, J., Pírková, E., & Wagner, B., 2021. "Reimaging Content Moderation and Safeguarding Fundamental Rights". *A Study on Community-Led Platforms.* [accessed 17 January 2025].

[231] Gorwa, R., Binns, R., and Katzenbach, C., 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance, *Big Data and Society,* 7(1) [accessed 29 May 2025].

[232] An additional 48 million decisions (31%) were partially automated, while only 11.8 million decisions (8%) involved no automation in the decision at all. See European Commission. DSA Transparency Database [accessed 29 May 2025].

unreported.[233] Several third-party providers supply technology to detect nudity and other sexual content,[234] as well as text-based content referencing self-harm, suicide, or eating disorders.[235]

9.90 Many providers currently use proactive technology to detect or support the detection of – and often, to take action on – target illegal content and/or content harmful to children. For example, Meta Platforms Inc. (Meta), TikTok, and Snap Inc. have stated in their transparency reports that they use proactive technology to detect pornographic content and/or pornographic content proxy.[236] Similarly, the use of proactive technology to detect or support the detection of fraud is also widely established across a variety of online services.[237] For further evidence on the accurate and effective deployment of existing proactive technology to detect or support the detection of target illegal content and/or content harmful to children, see 'Annex 13: Further evidence of relevant harms'.

9.91 We consider that the use of proactive technology to detect harms captured in this measure has the potential to result in increased detection of relevant harms, significantly reducing harm to users online, creating safer online experiences.

9.92 However, there are some limitations to the use of proactive technology in detecting or supporting the detection of the relevant harms. For example, proactive technology does not always deal well with nuance and context in the same way as humans. However, we mitigate this through the proactive technology criteria which are designed to ensure proactive technology is deployed in a way that ensures an appropriate balance between precision and recall, and that an appropriate proportion of content is reviewed by humans.

9.93 For further considerations relating to freedom of expression and privacy impacts, see the 'Rights assessment' section.

9.94 We consider that this proposed measure promotes the accurate and effective deployment of proactive technology through the use of proactive technology criteria, while including key safeguards to protect user rights and manage unintended consequences.

9.95 See 'Annex 13: Further evidence of relevant harms' for additional supporting evidence.

Impacts and costs

9.96 This section assesses direct and indirect impacts on service providers. Many aspects of this analysis are expected to apply similarly across different kinds of service providers and

[233] Thorn, 2023, Thorn's 2023 Impact Report [accessed 18 March 2025].

[234] SightEngine, Advanced Nudity Detection API. [accessed 13 June 2025] is an AI based service to determine the nudity content of images and videos; Amazon Rekognition. Content-Moderation: say they can detect and label explicit images and videos, that their image moderation returns a hierarchical list of labels which indicate specific categories of adult content. [accessed 14 June 2025].

[235] Moderation API, Self-harm model - Moderation API [accessed 13 June 2025].

[236] TikTok, 2024 Community Guidelines Enforcement Report [accessed 13 June 2025]. Snapchat Values, 2024, Snapchat Transparency Report | Snapchat Transparency [accessed 13 June 2025] Transparency Centre, 2019-2025, Community Standards Enforcement | Transparency Center [accessed 13 June 2025].

[237] Google, 2024. Written Evidence to the Home Affairs Select Committee [accessed 13 June 2025]; Meta, 2024. Testing New Ways to Combat Scams and Help Restore Access to Compromised Accounts [accessed 13 June 2025]; Twitch, 2023. Twitch State of Engineering 2023 [accessed 13 June 2025]; Bumble, Press Release: Community Report [accessed 13 June 2025]; LinkedIn, 2023, Augmenting our content moderation efforts through machine learning and dynamic content prioritization [accessed 13 June 2025].

content. Where relevant, we identify cases where costs may vary based on specific harms, media types, and other service characteristics, such as size.

9.97 The assessment in this section primarily focuses on the use of content identification technology. Costs may vary where providers implement proactive technology for user profiling or behavioural identification. We consider our assessment to be broad enough to capture an adequate range of cost scenarios in practice, while recognising that the actual implementation of the proposed measure may vary greatly in practice in different contexts and acknowledging that it is not possible to precisely assess every real-world scenario.[238]

Direct costs for service providers

9.98 The cost of implementing proactive technology is dependent on several factors, including the volume of content on the service to be scanned, the type of media to be scanned (such as text, image, audio, or video), the number, type and granularity of relevant harms, and the type of technology (for example, machine learning-based or rule/heuristic-based moderation technology).

9.99 Our analysis considers two separate scenarios: (1) a service that sources proactive technology from a third-party provider, and (2) a service that chooses to develop proactive technology in-house. This is a simplified approach for the purpose of our assessment. We expect that, in practice, some providers may use a combination of these approaches.

9.100 Our analysis shows that building proactive technology in-house would typically require investments and resources that may only be realistic for larger services. Therefore, our assessment of the costs of implementing the proposed measure using third-party software are particularly important in our analysis of whether the measure is proportionate for smaller services.

Costs of assessing and deploying proactive technology sourced from a third party

9.101 As part of step 1, providers would be expected to incur costs to identify relevant third-party software, assess it according to the proactive technology criteria, and to integrate and configure it for the needs of their service. Providers may rely, to some extent, on documentation from the third-party provider regarding how the software has been developed and tested and what safeguards are in place. However, we would expect services to carry out additional testing, quality assurance, and configuration (for example, of precision and recall levels) using their own data in an offline environment before deploying the software in a live environment.

9.102 Indicatively, this might require one to two software engineers working alongside one to two professionals from other occupations (such as analysts, product managers, and lawyers) for one to two months. For this, we estimate a cost of £10,000 to £90,000 using our standard salary assumptions. Providers of smaller services in scope of one kind of relevant harm and with one media type to scan are more likely to incur costs nearer the lower estimate. Larger,

[238] We note that existing content moderation measures in the Illegal Harms and Protection of Children codes already set more general expectations for providers to have a content moderation function, to set certain policies and to resource the function appropriately. This proposed measure sets more specific expectations with regard to the use of proactive technology to address specific harms. We assess the incremental impacts of the proposed measure, over and above the existing content moderation measures assumed to already be in place.

more complex services – and those in scope of multiple kinds of relevant harm and with multiple media types – are more likely to incur costs near or above the upper estimate.

9.103 Part of the costs associated with step 1 would be incurred even where a provider concludes it is not feasible to deploy proactive technology that meets the criteria.

9.104 As part of step 2, providers may incur additional internal costs associated with procuring the relevant technology (for example, making contractual arrangements and undertaking governance or budget processes), and deploying it. For smaller, less complex services, we expect such costs to be relatively low and to take less time. However, these costs may be substantial for larger services with more complex systems and technology stacks.

9.105 External fees paid to the third-party technology provider are likely to account for most of the cost under step 2. Fees tend to be based on the volume of content to be scanned. Additional fees, such as technical support fees, may apply in some cases.

9.106 Based on available evidence about third-party software with relevant capabilities for image or text detection, we assume that the unit cost for scanning one piece of content (for example, one image or 1,000 characters of text) to be in the range of £0.0005 to £0.005. Table 9.2 shows illustrative annual costs for some hypothetical services to scan new content posted. For example, if a service has 700,000 users – of which 20%[239] post one new piece of content every month on average – the annual cost of scanning new content could range from £900 to £9,000.[240] On a hypothetical larger service with 7,000,000 users – of which 60% post content once a month on average – the annual cost could be £26,000 to £260,000.[241]

[239] As an indication, we note that in Ministry of Internal Affairs and Communications, Information and Communication Policy Research, [accessed February 2025] has assumed, based on survey research, that on a service with 10 million users, 20% of users actively post content, and that they do so at least once a month. This assumption is used to set thresholds for certain rules to take down harmful content. We acknowledge that the volume of content posted on services will vary and have used some indicative percentage assumptions to capture plausible volumes of content posted by users on average. Ministry of Internal Affairs and Communications, 2023, Information and Communication Policy Research, https://www.soumu.go.jp/main_content/000976453.pdf, [accessed in February 2025].

[240] We expect the volume of content on a service to generally correlate with the number of monthly users. However, this relationship is unlikely to be exact and may vary significantly. On services where user engagement and posting frequency are particularly high, the volume of content is likely to be disproportionately greater. This trend would tend to drive up costs, however, this can be offset by two factors. Firstly, third-party providers often offer reduced pricing when higher volumes of content are scanned. Therefore, they would be eligible for volume discounts. Secondly, higher user engagement often correlates with greater revenue, which provides these services with a stronger ability to manage and absorb the associated expenses.

[241] We acknowledge the variability of these calculations across different services and have therefore considered different assumptions for illustrative purposes. Our aim is to capture the majority of real-world scenarios; however, we recognise that costs may still fall outside our projected ranges in certain cases, particularly for services with disproportionately high engagement or new content volume relative to their user base.

Table 9.2: Estimated cost of scanning new content using proactive technology[242]

Hypothetical service	Number of users	Percentage of users who post content once a month	Volume of new content (annual)	Estimated annual cost of scanning content	
				Lower bound (£0.0005 per unit)	Upper bound (£0.005 per unit)
Service A	700,000	20	1,700,000	£900	£9,000
Service B	700,000	40	3,400,000	£1,700	£17,000
Service C	7,000,000	40	34,000,000	£17,000	£170,000
Service D	7,000,000	60	50,400,000	£26,000	£260,000

9.107 Where pre-existing content is also scanned (as proposed for image-based CSAM, CSAM URLs and grooming), costs will depend on the volume of such content and could be significantly greater. For example, if a hypothetical service with a user base of 700,000 has five years' worth of pre-existing content available to UK users which has been generated at a constant rate, the cost of scanning pre-existing content could range from £4,000 to £42,000 for scanning all pre-existing text or image-based content on the service.[243]

9.108 We acknowledge that the volume of content is likely to be greater and the costs are likely to be higher on services with high user engagement and more frequent posting of content. However, it should be noted that third-party providers often offer discounted pricing for higher volumes, which mitigates cost impacts. In addition, services with larger user bases and higher user engagement may benefit from greater revenue-raising opportunities and may therefore be more able to manage these costs.

9.109 In any case, the technology used, and its cost may vary greatly depending on the context of a specific service. Costs are generally likely to be lower for some media types (such as text) and higher for media such as video, which could exceed the indicative estimates in Table 9.2 in some cases.[244] Costs may be higher where a service has multiple relevant risks and media types (such as text, images, video, and audio), and where the overall complexity of the service is greater (for example, where there are a larger number of different functionalities on which harmful or illegal content may be generated, shared, or encountered). However, some third-party software may be able to detect multiple relevant harms across different media types, meaning that costs may not increase in direct proportion to these factors.

[242] Numbers have been rounded to the nearest hundred or thousand where applicable for clarity and ease of interpretation.

[243] We recognise that platforms may regularly archive or delete old content, reducing the cost of scanning existing data compared to stated estimates here.

[244] For example, in the case of video content, it may be analysed by proactive technology by scanning images from multiple frames of the video, each one constituting a 'piece of content' in the sense of the indicative cost estimates of Table 9.2. Therefore, costs may be significantly higher for such content, particularly where videos are long. At the same time, the measure is not prescriptive, and providers have the flexibility to explore cost-effective implementation strategies, for instance in deciding any frame rate that is used for this purpose, provided that the criteria are met.

9.110 There may be further staff and non-staff costs on an ongoing basis. These may include costs of ensuring that the proposed recommendations are met with respect to monitoring the technology, storing inference logs for audit trails including user appeals and regulatory queries, reviewing the technology's performance, and managing feedback mechanisms. We have not quantified such costs because the extent to which they are incurred by regulated service providers or by third-party technology providers is likely to vary on a case-by-case basis.

Costs of developing and deploying proactive technology in-house

9.111 The cost of developing proactive technology in-house is likely to vary greatly depending on the context. Therefore, our analysis is indicative and considers the development of proactive technology designed to detect or support the detection of one particular kind of target illegal content and/or content harmful to children focusing on one media type (such as images). Costs may be significantly higher for some services, such as those that are particularly large and complex or have multiple relevant risks.

9.112 Costs are expected to include significant upfront investment and additional ongoing costs.

9.113 There would also be non-labour costs associated with other necessary inputs and resources such as training data, hardware and software.[245]

9.114 Purchasing training data may be costly where a service does not have relevant training data available. This cost is likely to vary greatly depending on circumstances (for example, the type of media and the type of harmful content). Indicatively, a large high-quality dataset could cost tens of thousands of pounds, or more. For example, the cost of acquiring a high-quality training dataset of 100,000 images could range between £100,000 and £300,000 assuming a price of £1 to £3 per image.[246] Larger services are generally more likely to hold relevant data internally but may still require significant time and resources to prepare datasets tailored to specific tasks.

9.115 Additional infrastructure costs would also vary depending on context. On a large service with complex existing systems, the infrastructure costs are likely to be substantial and could reach the same order of magnitude as labour costs.

9.116 Once the technology has been deployed, ongoing costs would involve regular maintenance and support. This includes system updates and bug fixes, as well as technology-related expenses like server management. Applying our standard cost assumptions, annual maintenance costs could range from £65,000 to £775,000 if they were equal to 25% of the initial set-up costs. Ongoing infrastructure costs could also be high, particularly for large and complex services. For example, a large technology company stated that its ongoing machine costs for training machine learning classifiers (for example, compute) and data storage reaches tens of millions of dollars per year.[247]

[245] We do not quantify these precisely as they are likely to vary significantly based on service providers' needs, but they are expected to be comparable to labour costs in many cases.

[246] According to the study by Dimensional Research, 2019, Artificial Intelligence and Machine Learning Projects Are Obstructed by Data Issues [accessed 13 June 2025], on average ML projects need around 100,000 data samples to perform well. However, this number might vary depend on the services' specific needs and structure.

[247] [✂] Response to our formal information request, February 2025.

Human moderation cost implications

9.117 Regardless of whether technology is developed in-house or sourced externally, its deployment may affect human moderation processes and the associated costs.

9.118 The impact will depend on the type of harm present on the service and on the provider's existing human moderation approach and processes prior to implementing the proposed measure.

9.119 On the one hand, proactive technology can result in an increased volume of content being detected. The proposed measure recommends that an appropriate proportion of this is reviewed by human moderators, potentially increasing the workload for human reviewers.

9.120 On the other hand, proactive technology can also be used to make automated moderation decisions, particularly in cases where there is a high likelihood of content being illegal or harmful to children. This approach could reduce the need for human review, but only for content that would have been flagged by human users in the absence of proactive technology. Otherwise, such content would remain undetected without the use of proactive technology. Any deterrent effect – where bad actors become aware that proactive technology is being used to tackle illegal content and/or content harmful to children – could also reduce human moderator workloads in some cases.

9.121 Consequently, the net effect is uncertain and may vary on a case-by-case basis. In general, it is more likely that the proposed measure will significantly increase human moderation costs where there are large volumes of target illegal content and/or content harmful to children present on the service. The measure recommends an appropriate balance between precision and recall,[248] allowing providers a degree of flexibility in determining this, which may help to mitigate cost impacts that could otherwise arise from dealing with disproportionate volumes of false positives.

9.122 Introducing, or adjusting, proactive technology could require changes to existing content moderation systems and processes covered by existing measures – for example, adjusting performance targets (ICU C4 and, where applicable, PCU C4) and/or prioritisation processes (ICU C5 and, where applicable, PCU C5).

Potential broader impacts on the market

9.123 The proposed measure could increase the cost of making content available to UK users. This could have knock-on impacts, such as a reduction in the pool of content, particularly on smaller services with more limited resources. In extreme cases, it could lead to services choosing to stop serving UK users altogether, although we consider this unlikely in most cases. Overall, the proposed measure could lead to some degree of reduction in the variety and volume of content available to UK users.

9.124 Certain types of services, such as those focused on video content, may incur higher costs compared to services that concentrate on media types (such as text) that are less costly to address using proactive technology. This differentiation could affect commercial decisions and competitive dynamics, although we expect that services that handle large volumes of

[248] High rates of incorrect detections can significantly increase operational costs, as manual review and escalation processes become necessary.

more complex content (such as video) would generally tend to be better resourced and therefore have greater capacity to implement the proposed measure.

9.125 However, there may also be positive impacts where users feel safer due to enhanced measures, leading to higher engagement and trust in the services provided.

Impacts on other stakeholders

9.126 In terms of impact on proactive technology providers, the proactive technology market appears likely to continue evolving at a steady pace. This measure may increase demand for proactive technology which could lead to more firms entering the market. In turn this could lead to more competition which would promote innovation and cost efficiency.

9.127 Consequently, there could be expanded business opportunities for these providers, potentially leading to an increase in market entrants. Greater competition in the sector may contribute to improved outcomes for service providers over time.

Rights assessment

9.128 In general, there is a substantial public interest in reducing the online prevalence and dissemination of illegal content and content harmful to children. It relates to the prevention of crime, the protection of children's health and morals, public safety, and the protection of the rights of others including victims and survivors of crime and users (including child users) of regulated services. The measures set out in this chapter are in pursuit of those objectives.

- The proposed measures (ICU C11 and PCU C9, ICU C12 and PCU C10) are intended to protect the rights of users which might otherwise be impacted by the online prevalence and dissemination of target illegal content and/or content harmful to children. These include: the right to respect for private and family life (Article 8 of the European Convention on Human Rights (ECHR)), which is a qualified right. In the context of this measure, all target illegal content and/or content harmful to children may impact the enjoyment of this right in the form of harm caused by the prevalence and/or dissemination of such content.
- The proposed measures are also intended to protect the right to life, a right which may be impacted by suicide, self-harm, and eating disorder content. Article 2 enshrines a positive obligation on public authorities to protect by law the right to life. It is an absolute right under Article 2.
- Article 3, another absolute right, imposes a positive obligation on public authorities to protect individuals from serious ill-treatment by others. Any reduction in child sexual abuse would protect children's Article 3 rights.

9.129 Where the recommendations in the proposed measures impact on human rights, that interference must be lawful, must pursue a legitimate social aim, and must not be disproportionate. Legitimate social aims include the protection of the rights and freedoms of others. In this case, freedom of expression (Article 10 of the ECHR) and other Article 8 considerations, such as how the measures impact on user privacy, may be impacted by the proposed measures. Adverse impacts on rights which may result from the proposed measures must be balanced against the substantial public interest in protecting people, especially children, from these harms.

9.130 The evidence of harm caused by the online prevalence and dissemination of target illegal content and/or content harmful to children, to the extent that it interferes with the rights set out above, is set out throughout this chapter (in particular in sections 'What risk does the use of proactive technology seek to address?' and in 'Annex 13: Further evidence of relevant harms'.

9.131 Further assessment of the impact on these rights, and the proportionality of that impact, is set out in the following paragraphs.

Freedom of expression and association

9.132 Article 10 of the ECHR sets out the right to freedom of expression. An interference with this right must be lawful, must pursue a legitimate social aim, and must not be disproportionate.

Potential interference with freedom of expression

9.133 The proposed measure may impact upon freedom of expression in several ways.

9.134 There is a risk that some content might be incorrectly identified (detected) as being target illegal content and/or content harmful to children. As a result, access to it might be restricted,[249] or it might be taken down. We recognise that as more content is detected and fed to content moderation, the volume of potentially incorrect content moderation decisions is likely to increase. Large scale occurrence of this on a service could amount to over-removal or over-restriction of content on that service which would have an adverse impact on users' rights.

9.135 Where a service has a higher tolerance for false positives, more content may be wrongly identified. This is because where a service has (or the proactive technology tool is set with) a higher tolerance for false positives, its systems are less designed to minimise the occurrences of incorrect detection of target illegal content and/or content harmful to children. We also acknowledge that PPC (pornographic, suicide, self-harm and eating disorder content) includes content that varies in terms of its severity and harm. This means that in some cases, detection may involve fine line judgements, which may result in benign content being incorrectly identified as content harmful to children.

9.136 The extent of false positives will depend on the service in question and the way in which it configures its proactive technology. The measure allows providers flexibility in this regard, including as to the balance between precision and recall (subject to certain factors set out earlier in this chapter). We recognise that this could lead to significant variation in impact on users' freedom of expression between services.

9.137 A user's freedom of expression might be impacted where incorrect detection results in the service taking action against them, suspending the user's account or shadow banning them. Shadow banning is when a user's content is downranked, removed from feeds, or similar by the service, sometimes without the user's knowledge. Services take these actions to limit the visibility of content as an alternative to banning or content removal. Restricting a user's

[249] We also recognise that in certain circumstances, adults' ability to access primary priority content that is harmful to children may be restricted under our existing content moderation measures. For example, PCU C2.3 recommends that where a provider's terms of service prohibit a kind of primary priority content, and the provider determines that content is that kind of primary priority content, it should swiftly take down the content (in which case it cannot be accessed by adult users in addition to child users) or take certain other steps where this is not currently technically feasible. Therefore, where a service provider incorrectly identifies content as primary priority content, this may impact adults' rights in addition to children's rights.

freedom of expression in this way may also cause stigmatisation. Stigmatisation would be particularly egregious where a user is incorrectly flagged as posting CSAM or grooming.

9.138 There is a risk that services may respond to the proposed measure by introducing more stringent terms of service, prohibiting more content than is required under the Act, or by our measures. This could lead to over-removal or over-restriction and could negatively impact users' freedom to post and share content.

9.139 As noted in the section describing broader impacts on the market, the proposed measure could also lead to some degree of reduction in the variety and volume of content available to UK users.

Proportionality of the potential interference

9.140 We consider that the potential interference with the right to freedom of expression is proportionate.

9.141 The proposed measure is designed in such a way that it should not have an undue adverse effect on freedom of expression. The proactive technology criteria we have developed stipulate that technology should strike an appropriate balance between precision and recall. This means that if available proactive technology generates a disproportionate number of false positives, then the proactive technology would not meet the criteria and services would not need to deploy it. We are recommending that services only deploy proactive technology if it meets the proactive technology criteria which relate to the technology's accuracy, effectiveness and lack of bias. This will mitigate the risk of over-detection (and by extension over-removal or over-restriction) because providers' systems will be designed to strike an appropriate balance between precision and recall (subject to considering various factors, including the proportion of detected content that is a false positive).

9.142 For the harms in scope of this measure, we have a higher degree of confidence that proactive technology that is accurate, effective and free from bias is likely to be available. Where proactive technology that meets the proactive technology criteria for accuracy, effectiveness and lack of bias in relation to a particular kind of target illegal content and/or content harmful to children is not available, the proposed measure recommends that the provider takes other steps (for example, keeping records of their assessment) in order to benefit from the 'safe harbour' (i.e. to be treated as complying with relevant safety duties). This safeguard will ensure that providers demonstrate the process they have been through to try to source proactive technology that meets the proactive technology criteria. It mitigates the risk that providers will implement technology that is inaccurate, ineffective or biased, which would be more likely to disproportionately interfere with users' freedom of expression.

9.143 Similar safeguards apply where a provider is using proactive technology supplied by a third party. Thus, providers will need to ensure that proactive technology sourced from third parties also meets the proactive technology criteria.

9.144 In relation to technical configuration, the proactive technology criteria also state that providers should configure the technology "so that its performance strikes an appropriate balance between precision and recall". In striking this balance, providers should take certain matters into account, namely:

a) the service's risk of relevant harm(s);

b) the proportion of detected content that is a false positive;

c) the effectiveness of the systems and/or processes used to identify false positives before moderation action is taken; and

d) in connection with CSAM or grooming, the importance of minimising the reporting of false positives to the NCA or a foreign agency.

9.145 Thus, it unlikely to be appropriate to configure proactive technology to have a higher tolerance for false positives unless there are other effective systems and/or processes to identify false positives as part of the content moderation process and therefore guard against over-removal or over-restriction of content. In that regard, the measure also recommends the review of an appropriate proportion of the outputs of the proactive technology during its deployment. Providers have flexibility in deciding what proportion of detected content is appropriate to review, taking into account the following factors:

a) The principle that the resource dedicated to the review of detected content should be proportionate to the degree of accuracy achieved by the technology and any associated systems and processes;

b) The principle that content with a higher likelihood of being a false positive should be prioritised for review; and

c) In the case of CSAM or grooming, the importance of minimising the reporting of false positives to the NCA or a foreign agency.

9.146 This will also reduce the potential impact of false positives detected by the technology.

9.147 It should be noted that this measure does not itself recommend the removal of detected content. Rather, it recommends that providers moderate detected content in accordance with existing content moderation measures (subject to human review of an appropriate proportion of detected content, as mentioned above).

9.148 Other measures act as safeguards to mitigate the risk of interference to freedom of expression. These include:

a) measures which enable users to complain or appeal if their content has been wrongly identified as illegal content [250] or content harmful to children;[251]

b) measures setting out appropriate action for complaints about proactive technology that results in content being taken down, restricted etc, which complaints are not appeals; [252]

c) a measure recommending that providers give information in their terms of service about proactive technology used for the purpose of complying with safety duties;[253]

d) a number of measures regarding content moderation systems and processes for large or multi-risk services, including:[254]

> setting internal content policies, rules, standards and guidelines around what regulated user-generated content is allowed on the service and what is not; how content that is harmful to children is to be dealt with on the service including whether

[250] Illegal Content U2U Code, ICU D1, D2, D7, D8, D9.

[251] Protection of Children U2U Code PCU D1, D2, D7, D8, D9.

[252] Illegal Content U2U Code ICU D11.

[253] Illegal Content U2U Code ICU G1.2 and Protection of Children U2U Code PCU G1.2.

[254] Illegal Content U2U Code ICU C3, C4, C5, C7, C8 and Protection of Children U2U Code PCU C2, C3, C4, C5, C7, and C8.

or not any kinds of content that is harmful to children are or are not allowed on the service; and how policies should be operationalised and enforced;[255]

- setting and recording performance targets for its content moderation function which should take account of the accuracy of decision making and should balance the need to take relevant content moderation action swiftly against the importance of making accurate moderation decisions;[256]
- setting a policy in respect of the prioritisation of content for review, which should include taking into account the desirability of minimising the number of users or child users who encounter a particular item of content; the severity of suspected harm to users or child users if they encounter the particular content (including whether the content is suspected to be priority illegal content, the risk assessment(s) of the service and the potential harm to children; whether the content is suspected to be primary priority content, priority content or non-designated content; and the likelihood that the content is illegal content and/or content harmful to children, including whether it has been reported by a trusted flagger);[257] and
- providing training and materials to voluntary and non-voluntary individuals working in content moderation to enable them to fulfil their role in moderating content.[258]

9.149 Providers have reputational and commercial incentives to try to limit the amount of content that is incorrectly moderated. These include meeting users' expectations and avoiding the cost of dealing with complaints and appeals. These commercial incentives should encourage providers to strike a reasonable balance in their implementation of the measure, including the technical configuration of their proactive technology and their use of human moderators.

9.150 Effective content moderation could positively impact enjoyment of the rights to freedom of expression and freedom of association. Safer spaces online make users feel more able to express themselves through creating content and other online activity. This may lead to higher engagement and trust in the services provided.

9.151 We are also recommending that providers should ensure feedback mechanisms are in place to maintain or improve its performance over time. This includes updating the technology with diverse and up-to-date datasets to reflect evolving trends or emerging types of harmful content and/or integrating ongoing feedback from users and individuals working in content moderation into its development, while managing the risk of introducing additional bias. This mitigates against the risk of the proactive technology generating biased outcomes.

9.152 We accept that services could adopt terms of service which prohibit a broader range of content than target illegal content and/or content harmful to children. If services take this approach, they may use proactive technology to detect or support the detection of a broader category of content than that for which they are at medium or high risk. This is a commercial matter for services and is not something recommended by this measure.

9.153 We have explained in the 'Benefits and effectiveness at addressing risks' section, why we consider that the proposed measure will be effective in detecting or supporting the detection of target illegal content and/or content harmful to children and therefore

[255] Illegal Content U2U Code ICU C3 and Protection of Children U2U Code PCU C2.

[256] Illegal Content U2U Code ICU C4 and Protection of Children U2U Code PCU C4.

[257] Illegal Content U2U Code ICU C5 and Protection of Children U2U Code PCU C5.

[258] Illegal Content U2U Code ICU C7-8 and Protection of Children U2U Code PCU C7-C8.

contributing to more effective content moderation online. We have explained that relying on complaints and human moderation would not be sufficient to detect such content at scale and set out information about the very serious nature of the harms in question. We have also limited the scope of the measure to those harms where there is sufficient evidence of existing proactive technology that meets the proactive technology criteria.

9.154 Having taken account of the nature and severity of the harms in question, the principles we have built into the measure to ensure that the technology used is sufficiently accurate, effective and lacking in bias, and the wider range of safeguards provided by other measures, we consider overall that the measure's potential interference to users' freedom of expression to be proportionate.

Privacy and data protection

9.155 Article 8 of the ECHR sets out the right to respect for private and family life. An interference with this right must be lawful, must pursue a legitimate social aim, and must not be disproportionate.

Potential interference with privacy

9.156 As a preliminary point, the focus of the proposed measure is publicly communicated content and metadata relating to content. Private communications, and metadata relating to them, are not in scope of the measure.

9.157 The impact on user privacy rights may vary depending on the harm type (for example, CSAM or fraud) and the proactive technology type (for example, whether they are tools which compare content against a database, machine learning tools, or others). We have drawn out these distinctions, where relevant, in our analysis in the following paragraphs.

9.158 In general, all content moderation, whether automated or human, will impact on the rights of individuals to privacy and their rights under data protection law.[259] The degree of interference will depend on the extent to which the nature of their affected content and/or communications is public or private – or, in other words, gives rise to a legitimate expectation of privacy.

9.159 Using proactive technology to detect or support the detection of content may involve automated processing at scale of inputs, which may often involve processing of personal data. Automated processing of personal data can lead to a number of possible data protection harms, such as loss of control of personal data, invisible processing or unwarranted surveillance.

9.160 The measure will involve the processing of 'historical' personal data in some circumstances[260] (for example, scanning personal data already on the service at the time the provider implements proactive technology) and the processing of 'new' personal data as it is provided to the service.

[259] We set out our rights assessments for our general content moderation measures in our December 2024 Statement: Volume 2 Service Design and User Choice, chapter 2. [accessed 14 June 2025]; and, April 2025 Statement: Volume 4 What should services do to mitigate the risks of online harms to children, chapter 14. [accessed 14 June 2025]

[260] The measure recommends this for services with the applicable risk level for image-based CSAM or CSAM URLs.

9.161 We are not recommending the use of a specific kind of proactive technology. As a result of these measures, services may decide to adopt tools which involve comparing content against a database.[261] In such cases, the way in which the underlying database is compiled could have an impact on user privacy. Wrongful inclusion in the database could negatively impact user privacy, as users are unlikely to expect their content to be used in this way.

9.162 There are additional considerations for machine learning algorithms. The technology may involve mass processing of data (including personal data as mentioned in paragraph 9.159) for automated classification and moderation of content. Unlike standard databases (such as those used by law enforcement which are focused on identifying illegal content), the content in these datasets may not be illegal content or content harmful to children. Instead, such content is used to identify and manage all kinds of content (such as that which violates a service's policies), as well as 'benign' content examples to improve the model's false positive and false negative rate.

9.163 We are also proposing that services may comply with the measure by adopting proactive technology which identifies behavioural signals, metadata or user activity that might indicate, for example, an elevated risk of grooming or fraudulent behaviour. This goes beyond analysis of the content itself.

9.164 Proactive technology that incorporates behavioural analysis may target users if there is reason to believe they are high risk.

9.165 Scanning and analysis of behavioural signals, metadata or user activity presents a potential interference with the user's right to privacy and their right to respect for family life. For example, the perpetration of grooming and fraud is inherently difficult to identify, and much of the activity often appears benign. In practice, this means that the detection of content of this nature will necessarily involve considering some benign activity. As a result, unwarranted surveillance of user activity or unwarranted profiling may occur and may amount to a potential interference with that user's right to privacy. Identification of users as being at risk of perpetrating these offences might also be damaging, depending on the action which services take as a result. Users who are falsely indicated as being at an elevated risk of grooming or fraudulent behaviour are likely to experience stigmatisation, which may impact their right to respect for family life.

9.166 The measure will recommend that an appropriate proportion of detected content is reviewed by human moderators. Human review of content that users have posted (or of content depicting or describing users) may negatively impact users' privacy.

9.167 In respect of CSAM, removal of CSAM content is vital for protecting the privacy rights of victims and survivors. To the extent that this measure recommends the use of proactive technology to detect or support the detection of CSAM, that needs to be balanced in the overall proportionality and rights assessment.

9.168 To the extent that CSAM or grooming is detected using proactive technology, and providers become aware of it through the content moderation process, providers are required to report the relevant user to a law enforcement authority. Section 66 of the Act (which is not yet in force but is likely to be by the time we issue our statement) sets out duties for providers of regulated user-to-user services to report to the NCA. Section 70(5)(b) also allows providers to report UK-linked CSEA elsewhere which is not otherwise reported.

[261] Though other types of proactive technology could be used, for example, machine learning technology.

These reports may include information about the user, and therefore this represents a risk to their privacy (particularly if the report is based on a false positive). A report based on a false positive which contained information about an identifiable individual would be a significant intrusion into their privacy, even if triage processes are in place to ensure that no further action is taken by law enforcement.

Proportionality of interference with privacy rights

9.169 We consider that the potential interferences with the right to privacy set out above are proportionate for the following reasons.

9.170 The measure only applies in relation to content communicated publicly and metadata relating to user-generated content communicated publicly (due to the restriction in Schedule 4 to the Act, para 13(4)). Therefore, the user sharing the content should have a reduced expectation of privacy in connection with that content or associated metadata.

9.171 Privacy impacts are limited by safeguards related to the technology's configuration. To the extent that the use of proactive technology may negatively impact users' privacy, as noted in the 'Explanation of the measure' section, we will recommend that services only use proactive technology if that technology meets the proactive technology criteria (which providers can assess for themselves). These include criteria related to the use of high-quality data, contextual testing and evaluation, and striking an appropriate balance between precision and recall. This will mitigate the risk of over-detection of content, and consequently the amount of content that providers will need humans to review an appropriate proportion of. In this way, the proactive technology criteria safeguard against disproportionate interferences with privacy.

9.172 Providers should ensure they act in accordance with data protection legislation and Information Commissioner's Office (ICO) guidance.[262]

9.173 UK GDPR places a specific restriction on making decisions based solely on automated processing of personal data, where the decision has legal or similarly significant effects. This is imposed by UK GDPR Article 22. So-called automated decision-making is only permitted where service providers have implemented certain safeguards for the data subject's rights, freedoms and legitimate interests. The ICO has provided guidance on these matters.[263]

9.174 To the extent that providers may use third-parties as part of the content moderation process, ICO guidance is clear that providers and third-party must identify their respective roles and obligations under data protection law and ensure that all requirements of that law are met.

9.175 Furthermore, to the extent that the measure recommends the use of behaviour identification or user profiling technology, providers should have regard to ICO guidance on automated decision-making and profiling.

9.176 We are satisfied that the measure can be implemented in accordance with data protection law. We consider that safeguards under data protection law, as explained in the various pieces of ICO guidance, will help ensure that the impact of automated processing on data protection and/or privacy rights is minimised.

[262] This includes ICO, 2024, ICO guidance on content moderation [accessed 13 June 2025].

[263] This includes ICO, 2024, What is content moderation and how does it use personal information? | ICO [accessed 13 June 2025].

9.177 As noted in paragraph 9.149, providers have reputational and commercial incentives to try to limit the amount of content that is wrongly actioned. We consider this will also limit any negative impact on user privacy derived from the risk of false positives.

9.178 The measure is likely to have a positive impact on the enjoyment of other rights of users. These rights are set out at the start of this ‘Rights Assessment’ section, and they include the right to respect to private and family life of victims and survivors, who might suffer an impact on their mental and physical health and social integration amongst other things. For example, CSAM is a violation of the privacy of the victims and survivors depicted, and removal of such content has a positive impact on those rights, as well as fundamental values and essential aspects of private life in relation to children, including health.

9.179 We consider that other measures act as safeguards for users’ privacy (including the protection of their personal data), such as measures related to complaints and appeals,[264] and the measure recommending that providers give information in their terms of service about proactive technology used for the purpose of complying with safety duties.[265] These measures tend to promote compliance with the data protection principles of (in particular) accuracy, fairness and transparency, and to assist users to exercise their rights under data protection legislation.

9.180 There is a risk that a user could be banned for sharing and/or posting CSAM due to a false positive identified by proactive technology, and that this could negatively impact their privacy. We consider this risk to be mitigated by (1) the provisions in the measure which recommend that services only adopt proactive technology if a service can build a system from it that is sufficiently accurate, effective and lacking in bias; (2) the provisions in the measure which recommend that services take account of the importance of minimising false positives being reported to the NCA; and (3) the fact that detected content will trigger the content moderation process, during which detected content will be reviewed; (4) the various safeguards within the user banning measure itself, which are discussed in Chapter 16 CSEA user banning.

Our provisional assessment

9.181 While we acknowledge the measure is likely to involve a degree of interference with users’ rights to freedom of expression, we are satisfied that the proposed measure pursues the aims of protecting the rights of others; preventing crime; and protecting health. It also corresponds to a pressing social need. The evidence of the harms caused by the online prevalence and dissemination of target illegal content and/or content harmful to children which justify the measure is robust. There is a substantial public interest in protecting users from these harms, especially children.

9.182 We are also satisfied that the mitigations and safeguards we have set out in this subsection are aimed at reducing incorrect detection as far as possible so that any interference with freedom of expression is as minimal as possible and proportionate to the aim of reducing the prevalence and dissemination of target illegal content and/or content harmful to children.

[264] Illegal Content U2U Code, ICU D1, D2, D7, D8, D9, and D11; and Protection of Children U2U Code PCU D1, D2, D5, D7, D8 and D9

[265] Illegal Content U2U Code, ICU G1.2(b); and Protection of Children U2U Code PCU G1.2 (b)

Which providers should implement this measure

9.183 Given the prevalence and severity of the relevant harms, and the essential role proactive technology plays in detecting and mitigating them at scale, we propose that this measure should apply to large user-to-user services that are medium or high risk and user-to-user services with more than 700,000 monthly UK users that are high risk of at least one of the following harms:[266]

- illegal harms: image-based CSAM, CSAM URLs, grooming, fraud and other financial services offences (fraud), encouraging or assisting suicide (suicide); and
- content harmful to children: PPC, which includes pornographic, suicide, self-harm and eating disorder content.

9.184 We propose that this measure should also apply to

- User-to-user services that are file-storage and file-sharing services of any size that are high-risk of unknown image-based CSAM.
- All user-to-user services which identify a high risk of grooming.

9.185 We expand upon our reasoning for including each of these types of service in scope of the measure below.

Large medium risk services

9.186 Given the size of these services, we consider that the illegal content that remains online and the content harmful to children that remains accessible to children, has the potential to reach large numbers of UK users and cause significant harm. Considering the volume of content on these services, we do not consider complaints, reporting and/or human moderation to be sufficient in detecting content suspected to be illegal content and/or content harmful to children.

9.187 We would therefore expect the application of this measure to large medium risk services to result in material reduction in relevant harm(s), thereby delivering significant benefits for UK users. Given the safeguards for freedom of expression and privacy considered in this measure and through other measures recommended in both the Illegal Content User-to-user Codes and the Protection of Children User-to-user Code, we consider the measure is proportionate for large medium or high-risk services.

9.188 While the measure would have significant costs, we provisionally consider the benefits of applying it to these services to be significant enough to justify these costs. We also anticipate that providers of this size will generally have the resources to be able to implement the measure.

Services with 700,000 or more monthly UK users who are high risk

9.189 While we understand that the cost may be a greater burden for services that are not large, given the severity of the relevant harms, we provisionally consider that it would be proportionate to apply the measure to high-risk services with 700,000 or more monthly UK users. Services of this size are still likely to have large volumes of content, and a high level of risk would indicate substantial harm to users that would not be addressed in the absence of proactive technology as recommended by this measure. We therefore consider that the

[266] In respect of the measures we are proposing to include in the Protection of Children User-to-user Code, these measures are limited to services likely to be accessed by children.

benefit of applying the measure to these services is likely to be large enough to justify the costs. This is particularly the case given that our analysis indicates that many of the costs are likely to scale with volume of content, which may often correlate with size of service.

File-storage and file-sharing services who are high risk of image-based CSAM

9.190 We are not proposing a user threshold on file-sharing services with high risk of image-based CSAM. Given the severity of harm CSAM causes and the prolific amounts of CSAM that even small file-sharing services can host, we consider it is proportionate to apply this measure to all such services. For further evidence and reasoning, see 'Automated content Moderation' in the December 2024 Statement.[267]

All user-to-user services which identify as high risk of grooming

9.191 We are proposing that services which are high risk for grooming regardless of size, should be in scope of the measure for the following reasons:

- While reliance on user reporting is an imperfect means of addressing all of the harms targeted by this measure, there are particular limitations of user reporting in relation to grooming content which could lead to very significant harm on small services. This is because content can often appear innocuous, and victims and survivors are extremely unlikely to report it for many reasons.[268] Where this type of content is reported, it is often many years after the abuse occurred.[269]
- We have also received specific intelligence from law enforcement of risks of grooming on sites that are likely to be high risk with fewer than 700,000 UK users.

9.192 Therefore, we consider that proactive technology should be deployed to detect or support the detection of this type of content on all services who identify as high risk for grooming, regardless of size.

Provisional conclusion

9.193 Given the harms this measure seeks to mitigate in respect of the relevant harms, as well as the risks of cumulative harm that user-to-user services pose to UK users, we provisionally consider this measure appropriate and proportionate to recommend for inclusion in the Illegal Content User-to-user Codes and the Protection of Children User-to-user Code. For the draft legal text for this measure, see ICU C11 and PCU C9, and ICU C12 and PCU C10 in Annex 7 Addenda to illegal content User-to-user Codes and Annex 9 Addenda to Children's Safety User-to-user Codes.

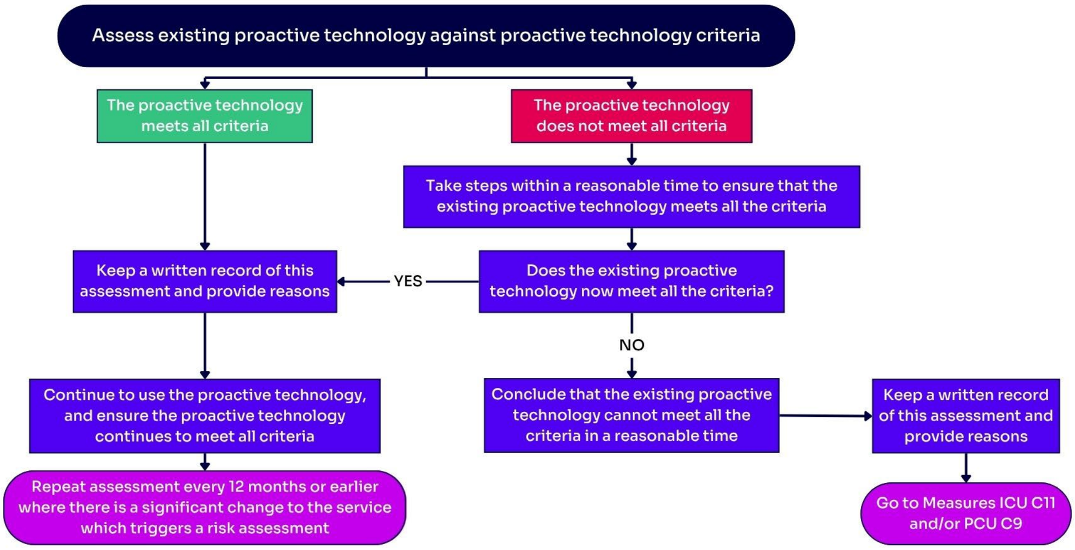
[267] Ofcom, 2024, Illegal Harms Statement: Volume 2 Service Design and User Choice, p.192 [accessed 13 June 2025].

[268] These reasons include shame, fear and the lack of recognition that a crime has occurred Quayle, E., Jonsson, L., Lööf, L., 2012. Online behaviour related to child sexual abuse. Interviews with affected young people. Council of the Baltic Sea States, Stockholm: ROBERT project. [accessed 22 May 2025]; Katz, C., Piller, S., Glucklich, T., & Matty, D. E., 2021. "Stop Waking the Dead": Internet Child Sexual Abuse and Perspectives on Its Disclosure. Journal of Interpersonal Violence, 36(9–10), NP5084–NP5104. [accessed 22 May 2025].

[269] Cited in Halvorsen, J. E. & Tvedt Solberg, E. & Hjelen Stige, S., 2020. "To say it out loud is to kill your own childhood". – An exploration of the first person perspective of barriers to disclosing child sexual abuse, Children and Youth Services Review, Elsevier, 113. [accessed 22 May 2025].

Measure ICU C12 and PCU C10: Assessing existing proactive technology for use to detect or support the detection of target illegal content and/or content harmful to children

Diagram 9.3: Flowchart visualising Measures ICU C12 and PCU C10


Assess existing proactive technology against proactive technology criteria
The proactive technology meets all criteria
The proactive technology does not meet all criteria
Take steps within a reasonable time to ensure that the existing proactive technology meets all the criteria
Does the existing proactive technology now meet all the criteria?
YES
Keep a written record of this assessment and provide reasons
NO
Continue to use the proactive technology, and ensure the proactive technology continues to meet all criteria
Conclude that the existing proactive technology cannot meet all the criteria in a reasonable time
Keep a written record of this assessment and provide reasons
Repeat assessment every 12 months or earlier where there is a significant change to the service which triggers a risk assessment
Go to Measures ICU C11 and/or PCU C9

Explanation of the measure

9.194 We propose this measure applies to providers who

- have already deployed proactive technology to detect or support the detection of target illegal content and/or content harmful to children before the implementation of ICU C11 and PCU C9; and/or
- have deployed proactive technology to comply with ICU C11 and PCU C9.

9.195 In summary, we propose that such providers should:

- Assess the proactive technology that they have already deployed to detect or support detection of target illegal and/or content harmful to children against the proactive technology criteria;
- if necessary, take steps within a reasonable time to ensure that the proactive technology meets all the proactive technology criteria or where the proactive technology cannot be changed to meet the criteria, to source new proactive technology; and
- record the outcome of their assessment.

9.196 Providers captured by this measure should update their assessment at a minimum of every 12 months, or earlier where there is a change to any aspect of the service's design or operation which would amount to a significant change and therefore trigger a risk assessment under section 9(4) of the Act.[270]

9.197 There may be rare circumstances where a provider concludes that it is not feasible for them to make changes to existing proactive technology so that it meets the proactive technology criteria, or source new proactive technology which meets the proactive technology criteria. In this instance, we consider it is a commercial decision for the provider as to whether they continue to deploy the existing proactive technology. Providers should record the outcome of their assessment. See 'Where providers cannot deploy proactive technology' in ICU C11 and PCU C9.

Step 1: Assess proactive technology which providers have deployed against the proactive technology criteria

9.198 As explained in section 'What is proactive technology?', we have developed proactive technology criteria that we propose services should consider when complying with these measures.

9.199 Further guidance on the proactive technology criteria, including how providers may assess technology against them, is set out in our supporting draft guidance.[271]

9.200 Where the outcome of the assessment is that the proactive technology deployed by the provider meets the proactive technology criteria, providers should record this outcome. In this instance, providers do not need to make changes to their proactive technology, on the basis that it is sufficiently accurate, effective and lacking in bias. See 'Providers should record the outcome of their assessment'.

9.201 We propose that providers, including those who deploy proactive technology once they have complied with ICU C11 and PCU C9, should update their assessment at a minimum of every 12 months, or earlier where there is a change to any aspect of the service's design or operation which would amount to a significant change and therefore trigger a risk assessment under section 9(4) of the Act.[272]

Step 2: If necessary, take steps within a reasonable time to ensure that the proactive technology meets all the proactive technology criteria.

9.202 This step applies in instances where the outcome of a provider's assessment is that the proactive technology does not meet all the proactive technology criteria.

9.203 Having identified the criteria that the proactive technology does not meet, providers should take steps within a reasonable time to ensure that the proactive technology meets all the criteria. This could involve making changes to the existing technology within a reasonable timeframe and/or sourcing new technology.

[270] See, Ofcom, 2024, Risk Assessment Guidance and Risk Profiles and Ofcom, 2025, Children's Risk Assessment Guidance and Children's Risk Profiles [accessed 13 June 2025], for when risk assessment(s) should be reviewed and updated.

[271] Proactive Technology Draft Guidance

[272] See, Ofcom, 2024, Risk Assessment Guidance and Risk Profiles and Ofcom, 2025, Children's Risk Assessment Guidance and Children's Risk Profiles [accessed 13 June 2025] for when risk assessment(s) should be reviewed and updated.

9.204 While providers are taking steps to ensure that the proactive technology meets all the criteria, we encourage providers to consider ways to ensure the continued safety of their users.

9.205 Providers may refer to Ofcom's Proactive Technology Draft Guidance[273] for further information and examples of how providers can make changes within a reasonable time to ensure that the proactive technology meets all the proactive technology criteria.

9.206 Where a provider concludes that their proactive technology cannot meet the proactive technology criteria within a reasonable time, providers should assess whether different proactive technology is available, is technically feasible to implement, and meets the criteria. See 'Sourcing proactive technology' in measure ICU C11 and PCU C9.

Step 3: Providers should record the outcome of their assessment

9.207 When a provider has completed their assessment of the proactive technology against the proactive technology criteria, they should record the outcome. This includes where:

a) the proactive technology meets the criteria;
b) a provider initially considered that the technology did not meet the criteria but has taken steps within a reasonable time to ensure the proactive technology meets all the criteria; or
c) the proactive technology does not meet the criteria.

9.208 The record should include:

- which of the proactive technology criteria that their proactive technology or combination of technologies did not meet (where applicable).
- the steps that the provider took to ensure that the proactive technology met all the proactive technology criteria (where applicable); and
- evidence that it now meets the proactive technology criteria (where applicable).

Benefits and effectiveness at addressing risks

9.209 We consider that this measure will help providers improve the accuracy and effectiveness of the proactive technology that providers are deploying to detect the relevant harms on their service.

9.210 Evidence from stakeholders highlights the importance of evaluating proactive technology to ensure it remains sufficiently accurate, effective and free from bias during deployment.[274] For example, some providers will rely on active learning cycles, scheduled or ad-hoc updates to ensure that their proactive technology is consistently meeting their performance

[273] Proactive Technology Draft Guidance
[274] Responses to our formal information request (Jan 2025) indicate that many providers have either internal risk intelligence teams including LinkedIn, [✂], Reddit, and Pinterest, and/or rely on third party providers to monitor trends and behaviours that may impact the effectiveness of their proactive technology over time. For example, Yubo explained that if they see an unexpected spike in appeals or user reports, suggesting that their proactive technology is not performing as intended, or there is a new/emerging threat that it has not accounted for, they will review and update their technology; this is one of several indicators they use to assess performance. One large service provider [✂], stated that they track their system's overall safety targets on a weekly basis in order to ensure their proactive technology is consistently meeting performance targets. Response to Ofcom's formal request for information, February 2025.

targets, reducing the amount of illegal content and/or content harmful to children for users. This may involve incorporating new data, adapting the technology to evolving patterns, or emerging threats and addressing technology 'concept drift'[275] when model performance changes.[276] Providers and third-party providers also stressed that improving the performance of a model is an on-going process, with evidence of long-term initiatives obtaining the best performance levels.[277]

9.211 By ensuring that proactive technology meets the criteria, we consider that this measure will increase the amount of illegal content and/or content harmful to children that service providers detect and in turn action on their service. We consider that this will deliver significant benefits to UK users, particularly children. We consider that the measure will also reduce some of the adverse impacts on freedom of expression and privacy which arise from the deployment of less accurate and/or effective proactive technology or proactive technology that may be biased.

9.212 We understand that proactive technologies differ in their accuracy, effectiveness and susceptibility to bias depending on the type of content and harm they are designed to detect or support the detection of. Therefore, we do not consider it to be appropriate at this time to prescribe specific thresholds for proactive technology. This is because:

- we understand that these can vary according to the harm, content, and service type;
- there are a lack of consistent, comparable testing methodologies across industry
- there is a lack of industry-defined performance standards for proactive technology; and
- setting numerical thresholds may have an adverse impact on freedom of expression, innovation and market growth.

9.213 We consider that by recommending that providers should update their assessment at a minimum of every 12 months, or earlier where there is a change to any aspect of the service's design or operation which would amount to a significant change and therefore trigger a risk assessment under section 9(4) of the Act, there is a potential to identify new and evolving harms or how the content manifests on a service that proactive technology should be deployed to detect the relevant harm(s).

Impacts and costs

9.214 This proposed measure applies to providers that already implement some relevant proactive technology. In this case, the implications for each provider in scope would depend on whether their approach meets the expectations set by this proposed measure or requires additional work to fulfil. As such, they are expected to vary significantly on a case-

[275] Evidently.ai, 2025, What is concept drift in ML, and how to detect and address it [accessed 13 June 2025.]

[276] A research report by Winder.ai explained how providers use metrics for training and governance to track the performance of their proactive technology. Governance metrics are often evaluated by trust and safety teams to help them decide whether a model is "good enough" to be deployed, while key performance indicators (KPIs) are often used to demonstrate that models are achieving their long-term goals. Winder.ai on behalf of Ofcom, 2023, Automated Content Classification (ACC) Systems, p.21 [accessed 13 June 2025]

[277] Winder.ai on behalf of Ofcom, 2023, Automated Content Classification (ACC) Systems, p.21 [accessed 13 June 2025.]

by-case basis and cannot be quantified precisely. In this section, we assess these impacts qualitatively and where possible, provide an indication of potential magnitude.

9.215 If a provider's existing technology already meets all the criteria and other relevant requirements, confirming and documenting this should entail only limited cost. Indicatively, the assessment may require several days or weeks of work for a single employee in the case of a small service, or a few weeks for a larger, more complex service.

9.216 For providers whose proactive technology does not meet all the proactive technology criteria, further investment may be needed. The scale of these costs will depend on the extent of required modifications. For example:

a) If existing technology meets most of the criteria, but only partially fulfils certain other criteria, additional testing or targeted adjustments may be necessary, potentially requiring a few weeks of work.
b) If existing technology and related systems or processes require more fundamental testing or adjustments to meet the criteria, then the costs may be significantly higher, potentially involving multiple months of work.
c) In cases where the most extensive changes are needed, including instances when a provider concludes that meeting the criteria requires developing or sourcing entirely new technology, the costs are likely to be the highest. As an upper bound, costs in this case could reach the same level as those assessed under proposed measure ICU C11 and PCU C9.

Rights assessment

9.217 This proposed measure is intended to protect the rights of users which might otherwise be impacted by the online prevalence and dissemination of target illegal content and/or content harmful to children. The proactive technology criteria used for the purpose of this assessment are the same as those used for the purpose of ICU C11 and PCU C9.

9.218 We expect that once ICU C11 and PCU C9 and ICU C12 and PCU C10 are implemented, they will impact users' human rights in similar ways. This is because both measures recommend that providers deploy proactive technology so that it continues to meet the proactive technology criteria. For this reason, the ways in which ICU C11 and PCU C9 are likely to impact human rights (as set out above) are likely to be relevant to ICU C12 and PCU C10 as well.

9.219 However, the extent to which ICU C12 and PCU C10 are likely to impact those rights will be different. ICU C12 and PCU C10 involve reviewing existing proactive technology and improving its performance to meet the proactive technology criteria. We do not expect ICU C12 and PCU C10 to significantly widen the circumstances in which providers use proactive technology. Rather we expect them to improve the quality of existing proactive technology in use. In respect of the rights set out above,[278] we consider that ICU C12 and PCU C10 are likely to result in proactive technology being used in a more rights respecting way than it is today. In other words, we expect this measure to reduce the extent to which the use of proactive technology may interfere with rights when compared with the status quo. This is not the case for ICU C11 and PCU C9, because introducing proactive technology where none was used before necessarily increases the potential for adverse impacts on users' rights.

[278] See 'Rights assessment' under ICU C11 and PCU C9.

Which providers should implement this measure

9.220 The proposed measure only applies where providers have already sourced or developed and deployed a proactive technology for the relevant harm(s).

9.221 As with ICU C11 and PCU C9, given the prevalence and severity of the relevant harms, and the essential role proactive technology plays in mitigating them, we propose that this measure should apply to large user-to-user services that are medium risk and user-to-user services with more than 700,0000 monthly UK users that are high risk of at least one of the following harms: [279]

- illegal harms: image-based CSAM, CSAM URLs, grooming, fraud and other financial services offences (fraud), encouraging or assisting suicide (suicide); and
- content harmful to children: PPC, including pornographic, suicide, self-harm and eating disorder content.

9.222 We propose that this measure should also apply to

- User-to-user services that are file-storage and file-sharing services of any size that are high-risk of image-based CSAM.
- All user-to-user services which identify a high risk of grooming.

9.223 For our reasoning as to why we propose this measure should apply to these specific services, see Measure ICU C11 and PCU C9, 'Who should implement this measure'.

Provisional conclusion

9.224 Given the harms this measure seeks to mitigate in respect to the relevant harms, as well as the risks of cumulative harm user-to-user services pose to UK users, we provisionally consider this measure appropriate and proportionate to recommend for inclusion in the Illegal Content User-to-user Codes and the Protection of Children User-to-user Code. For the draft legal text for this measure, see ICU C12 and PCU C10 in Annex 7 Addenda to illegal content User-to-User codes and Annex 9 Addenda to Children's Safety User-to-user Code.

Other options considered

9.225 We considered other options when developing these measures, including whether to extend the measures to a wider range of harms and services. We considered:

- proactive technology for other priority illegal harms;
- proactive technology for PC and NDC;
- proactive technology for search services; and
- additional recommendations for user-to-user services unable to deploy proactive technology.

9.226 We intend to continue building our evidence base, and we welcome evidence from stakeholders on the use of proactive technology for these options, including how accuracy,

[279] In respect of the measures, we are proposing to include in the Protection of Children User-to-user Code, these measures are limited to services likely to be accessed by children.

effectiveness and lack of bias are assessed, how trade-offs are managed, and how rights impacts can be mitigated.

Proactive technology for all priority illegal harms

9.227 We considered extending the measures to apply to all priority illegal harms:

- hate
- harassment, stalking, threats and abuse,
- controlling or coercive behaviour,
- drugs and psychoactive substances;
- firearms and other weapons;
- unlawful immigration;
- human trafficking;
- sexual exploitation of adults;
- extreme pornography;
- image-based adult sexual offences;
- proceeds of crime;
- foreign interference; and
- animal cruelty.

9.228 While we are aware that there is proactive technology to detect some of these harms, such as terrorism and intimate image abuse where we are proposing separate measures (Measures ICU C13 and ICU C14 in chapters 11 and 12 respectively) we did not consider it was proportionate to extend the measure to these additional harms.

9.229 There are more than 130 offences in scope of the Act (and a wide range of content that is harmful to children). To ensure that these measures are proportionate, we have limited the harms for which service providers should assess the availability of proactive technology, its technical feasibility and whether it meets the proactive technology criteria.

9.230 We have prioritised a subset of harms for these measures. These are harms which align closely with our eight targets for immediate action,[280] and where we have a higher degree of confidence that proactive technology that is accurate, effective and free from bias is likely to be available.

9.231 If we proceed to implement this measure, in due course we will review whether to expand the harms in scope of the measure in light of evidence following its implementation.

[280] For our eight targets for immediate action see Ofcom, 2024, Implementing the Online Safety Act: progress update p.8 [accessed 13 June 2025.]

Proactive technology for PC and NDC

9.232 We considered extending the measures to apply to PC and NDC that is harmful to children. This includes hate and abuse content, violent content, bullying content, and content promoting dangerous challenges.

9.233 Under the Act, providers have a duty to protect children in age groups judged to be at risk of harm from encountering this content. We set out examples of how providers could do this in our measures on reviewing, assessing, and taking action on content (PCU C1 and PCU C2) in the Protection of Children User-to-user Code.

9.234 While we are aware that there is proactive technology to detect some of these harms, such as hate and abuse and violent content, we did not consider we had a sufficient evidence base to propose a measure on sourcing and/or developing and deploying proactive technology for PC and NDC. We welcome further evidence on the accuracy and effectiveness of technologies for detecting PC and NDC, as well as how impacts on rights would be managed.

Proactive technology for search services

9.235 We considered extending this measure to suggest that proactive technology is used to detect relevant harms on search services.

9.236 We understand that some technology is used by search services, including some to tackle fraudulent content, as well as basic deindexing/downranking measures for illegal content and/or content harmful to children.

9.237 However, we are not proposing to extend the measure for three reasons. First, we have previously received feedback from stakeholders that the complexity of automated moderation of organic search content means that we cannot transfer principles developed for user-to-user services across to search services. Second, we have limited information on how moderation is applied to each ‘layer’ of search (crawling, indexing, serving results) that would enable us to develop appropriate principles for the measure. Third, we do not currently consider our evidence base to be robust enough to recommend all search services deploy this technology, at this time.

Additional recommendations for user-to-user services unable to deploy proactive technology

9.238 We considered an additional recommendation for providers to invest resource and effort into the development of proactive technology for high-risk relevant harms if they judged they were unable to source proactive technology for these harms for any reason as part of this measure.

9.239 We had originally considered proposing this requirement to increase the number of providers overall that would eventually deploy proactive technology on their service, despite proactive technology being unavailable at the time of recommendation. However, we do not currently consider this to be proportionate, especially for smaller services (for whom this would likely be a considerable investment). It may also be an inefficient use of a service’s resources, which could be deployed elsewhere (for example, for compliance with other areas of the Codes). However, this should not discourage providers from carrying out research and development efforts into new proactive technology if they have the means to do so.

10. Amendments to Illegal Content Judgements Guidance for child sexual abuse material

Summary

We know that perpetrators use messaging, group chats or forums to facilitate the commission of CSAM offences. In some cases, they use the name, icon or bio / description to indicate to other perpetrators that CSAM content can be found within. For some service providers it may be technically feasible to review those pieces of information but not the content of the messages, group chat or forum. We are proposing to amend our Illegal Content Judgements Guidance (ICJG) to guide providers about how to make illegal content judgements in these circumstances. This chapter sets out what the amendments are, why we propose to make them and why we consider the potential interference with users' freedom of expression to be proportionate.

To read the amendment in full, please see Annex 11: Proposed amendments to the Illegal Content Judgements Guidance.

Consultation questions

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

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

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

CSAM in messaging, group chats and forums

10.1 Our evidence shows that perpetrators use messages, group chats and forums to share CSAM, as well as ideas, advice and instructions about abusing children.[281] For example, a perpetrator may commit a CSAM priority offence by showing and/or distributing an indecent image or film of a child, linking or directing a user to CSAM or publishing an obscene article by messaging another potential perpetrator. A perpetrator may also commit a priority CSAM offence by coming into the possession of an indecent/prohibited image of a child or a paedophile manual through messaging with other perpetrators. A perpetrator may also use messaging, group chats and discussions on forums to encourage or assist another user to commit a priority CSAM offence, for example by encouraging the making of an indecent image or film of a child.

[281] Ofcom, 2024. Illegal Harms Register of Risks, paragraph 2B.56 [accessed 14 June 2025].

Perpetrators often use messaging, group chats and forums in combination with associated content to signpost the existence of CSAM

10.2 Perpetrators use messaging, group chats or forums to facilitate the commission of CSAM offences without detection. This is particularly the case in respect of offences of distribution, or encouragement to distribute illegal images of children. Offending can take place in a chat between two perpetrators, or more widely in a group chat or forum, and the image, bio or heading is sometimes used as a means to indicate to other perpetrators that CSAM content is contained within. We are aware that perpetrators may use "phrases, keywords, or other hints that signpost to illegal content", which "allow abusers to identify and form networks with each other" to commit or facilitate CSEA offences.[282] This is known as "breadcrumbing".[283]

10.3 In many cases the content within messages, group chats or forums will be privately communicated, and the connected content which is used to indicate the presence of CSAM will be publicly communicated, although this depends on the circumstances.[284]

It may be technically feasible for providers to review the name, icon or bio / description, but not the contents of messages, group chats and forums

10.4 In some cases, it may be technically feasible for providers to review the name, icon or bio / description, but not the content within messages, group chats and forums used by perpetrators.[285]

Inferring the presence of CSAM content in messaging, group chats and forums where it is not technically feasible for the service to review the content

10.5 Service providers may become aware (through third party databases, the use of technology as part of content moderation, or through police or user reports) of content which:

a) it is technically feasible for them to review, and
b) indicates the presence of CSAM content within messages, group chats or forums which it is not technically feasible for the service to review.

[282] NSPCC, 2023. 82% rise in online grooming crimes against children in the last 5 years. [accessed 06 May 2025].

[283] NSPCC, 2023. 82% rise in online grooming crimes against children in the last 5 years. [accessed 06 May 2025].

[284] See our Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act [accessed 17 June 2025] for further detail. We note that our proactive technology measures only recommend the analysis of content which is publicly-communicated: Schedule 4 to the Act, paragraph 13(4). Where the name, icon or bio / description is used to indicate to other perpetrators that CSAM content can be found within, it is more likely to be publicly-communicated (and therefore capable of detection under our proactive technology measures), although this will depend on the circumstances.

[285] The Act is clear that a provider must be able to consider complaints, which means a provider needs to be able to review content complained of. But in the case of private messaging, group chats and forums the participants themselves may be unlikely to complain. Concerns are more likely to be raised by non-participants whose complaints cannot necessarily be linked to specific items of content on the thread.

10.6 We are proposing to amend the CSAM chapter of our ICJG to give providers guidance about how to make illegal content judgements in these circumstances. The Act requires that service providers make judgements about whether content is illegal content based on a reasonable inference. The Act states that content will be illegal content where there are reasonable grounds to infer that: a) the conduct element of a relevant offence is present or satisfied; b) the state of mind element of that same offence is present or satisfied; and c) there are no reasonable grounds to infer that a relevant defence is present or satisfied.

10.7 We are proposing to address two different scenarios in the ICJG, as detailed below.

The first scenario: inferring illegality based solely on content which the provider has reviewed

10.8 We consider that content that it is technically feasible for providers to review, or complained-about content, may on its own give providers reasonable grounds to infer that other content within messages, group chats or forums (which it is not technically feasible for them to review all or part of) is illegal CSAM content, absent clear evidence to the contrary.

10.9 We consider this to be the case when:

a) the image or icon for the messages, group chat or forum is CSAM;
b) the image or icon for the messages, group chat or forum is a still from a known CSAM video; or
c) the bio or heading for the message, group chat or forum explicitly indicates that the contents are CSAM.

10.10 This is because we know that perpetrators use messaging, group chats or forums to facilitate the commission of CSAM offences without detection. As set out above, this is particularly the case in respect of offences of distribution, or encouragement to distribute illegal images of children, but several other CSAM offences are also committed in this way. Offending can take place in a chat between two perpetrators, or more widely in a group chat or forum. We are not aware of any reason for using CSAM, a still from a known CSAM video, or a bio or heading that explicitly indicates that the contents are CSAM, other than to signal to other perpetrators that the purpose of the chat is to commit one or more further CSAM offences. Therefore, where such content is used, and absent clear evidence to the contrary, we think this provides reasonable grounds to infer that the contents of the messages, group chat or forum contain CSAM and their users are aware of and intend this.

The second scenario: inferring illegality based on content which the provider has reviewed, in combination with other information

10.11 The second scenario is where the bio or heading relating to a message, group chat or forum indicates that the content of the chat is CSAM through coded terms strongly associated with CSAM. In such cases, and where it is technically feasible for providers to review such content, they should consider this as reasonably available information for inferring whether the content within the message, group chat or forum (which it is not technically feasible for them to review all or part of) is CSAM. However, providers should consider this content in combination with other available information, such as user complaints or police reports.

10.12 The reason for this is because perpetrators may use phrases, keywords or other hints to indicate or signpost the presence of CSAM content within messages, group chats or forums. These phrases, keywords or other hints may not themselves be illegal and may not explicitly indicate that the contents of the messages, chat group or forum contains CSAM. However, when combined with the other information referred to above, this may provide a sufficient basis for a provider to reasonably infer that the contents of the messages, chat group or forum are CSAM and their users are aware of and intend this.

Rights Assessment

10.13 We recognise that the approach set out above may amount to a significant interference with users' rights to freedom of expression.

10.14 Generally, providers review content to make an illegal content judgement about it before taking it down. Reviewing the content (in line with our ICJG) mitigates against the risk that legal content is taken down. The approach set out in the scenarios above involves taking content down without reviewing it. There is, therefore, a greater potential for the judgement to be incorrect and this could have a significant impact because it affects all the content on the message thread, chat or forum. However, as set out above, we are only proposing that providers should make this judgement in circumstances where they have clear evidence that the very purpose of the message thread, chat or forum is to commit CSAM offences.

10.15 In addition, we recognise that the approach involves an illegal content judgement about one piece of content (the content that it is technically feasible for the provider to review) potentially resulting in the take down of two sets of content (the content that has been reviewed; and the other content which it is not technically feasible to review). This means that if the provider makes an incorrect illegal content judgement about the first mentioned content, the consequential impact on freedom of expression would be greater. We consider that there are several factors which reduce the risk that providers may incorrectly identify the content they have reviewed as CSAM. Providers should follow the ICJG when considering whether content is CSAM. We have already introduced a measure in our Illegal Content User-to-user Codes, which recommends that certain providers use hash matching to detect and take down CSAM content. In this consultation we are also recommending that certain providers use proactive technology to detect CSAM content, and that detected content subsequently goes through the provider's content moderation. Both measures include safeguards to mitigate against the incorrect identification of CSAM.

10.16 We recognise that the risk of incorrect removal of content which it is not technically feasible to review is different in each of the two scenarios referred to above. In the ICJG, we recommend different approaches to making illegal content judgements in the two scenarios to reflect this.

10.17 We have also considered the privacy rights of users. There is a risk that where an illegal content judgement is wrong, it may amount to the creation of incorrect private information about the users of the message, group chat or forum. It may also interfere with a users' rights to private and family life to the extent that it interrupts private communications. However, the proposed amendments address a very serious harm. They apply in limited circumstances, and we are satisfied that the risk of incorrect removal is mitigated as set out above.

10.18 We are also mindful that:

a) where it is not technically feasible for providers to review the content in messages, group chats or forums; and

b) where the private content of messages, group chats or forums cannot be detected by proactive technology which we have recommended that service providers use;

c) there is a significant risk that perpetrators exploit this to distribute CSAM.

10.19 Taking the above into account, we consider that when providers follow the approach set out in our proposed amendments, the interference with users' rights to freedom of expression and privacy is likely to be proportionate to the significant benefits of reducing the prevalence and distribution of CSAM online, particularly in circumstances where it is not technically feasible for providers to review the content in messages, group chats or forums.

11. Perceptual hash matching for intimate image abuse

Summary

The Online Safety Act 2023 creates offences of sharing or threatening to share intimate images without consent. We propose that certain service providers use an automated tool called hash matching to detect intimate images that have been shared without consent, so they can be removed.

Intimate image abuse has very significant adverse personal, professional and psychological impacts on survivors and victims. This includes a constant threat from intimate images which are not removed or might resurface online. Intimate image abuse is widespread – evidence suggests it is increasing in prevalence and the proliferation of generative artificial intelligence (GenAI) tools is causing a further increase.

We understand that several service providers already use hash matching to tackle intimate images shared without consent, and the evidence suggests that it is an effective tool. We therefore consider that more widespread use of hash matching would reduce the volume of image-based intimate image abuse being shared and viewed online. We consider that this would materially reduce the significant negative impacts on the lives of survivors and victims.

Our proposals

Number	Proposed measure	Who should implement this
ICU C14	Providers use perceptual hash matching to detect image-based intimate image abuse content so it can be removed.	Providers of user-to-user services which are high risk of intimate image abuse and either: • whose principal purpose is the hosting or dissemination of regulated pornographic content[286]; or • are file-sharing and file-storage services; or • have more than 700,000 monthly active UK users. Providers of large[287] user-to-user services that are medium risk for intimate image abuse.
ICS C8	Providers use perceptual hash matching to detect image-based intimate image abuse content so it can be moderated.	Providers of large[288] general search services.

[286] Regulated pornographic content is pornographic content which excludes content which consists only of text or text accompanied by one of more of the following: identifying content consisting only of text, identifying content which is not pornographic, a GIF which is not pornographic, or an emoji or other symbol.
[287] Large refers to any user-to-user service with more than 7,000,000 monthly UK users.
[288] Large refers to any search service with more than 7,000,000 monthly UK users.

Consultation questions

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

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

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

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

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

Introduction

11.1 In this chapter, we set out our proposal for certain service providers to use an automated content and search moderation technology called perceptual hash matching to detect intimate images and videos which are non-consensually shared online. We refer to these images and videos as image-based intimate image abuse (IIA) content.

11.2 This chapter explains our proposals and is structured into four sections:

i. In the first section, we describe the harm caused by intimate image abuse to individuals, including survivors and victims.

ii. In the second section, we outline the proposed measure that service providers would be expected to implement (we detail the measure further in 14), including how the proposal interacts and complements other existing measures in the Illegal Content Codes of Practice (Illegal Content Codes).

iii. In the third section, we assess the impact of the proposed measure, including its effectiveness at reducing intimate image abuse and the benefits it would bring; the costs of the measure; and their impact on the fundamental rights of individuals.

iv. Finally, in the fourth section we explain why, on balance, we consider the proposed measure to be proportionate and discuss the service providers that we propose the measure should apply to.

The harm caused by intimate image abuse

11.3 The Online Safety Act 2023 (the Act) establishes priority offences which relate to sharing, or threatening to share, intimate images (photographs or films) without consent. In our Illegal Content Judgements Guidance, we explain how service providers should identify and take down such content, including in respect to the sharing of intimate images as reposts and re-shares where the image has been posted or shared before without consent.[289] [290]

[289] For further detail, see Section 10. Image-based adult sexual offences of the Illegal Content Judgements Guidance. [accessed 2 April 2025].

[290] We understand that the Government intends to introduce new offences relating to intimate image abuse. Specifically, the Crime and Policing Bill as currently drafted contains the new offences of taking or recording intimate photograph or film without consent and installing equipment to enable taking or recording of intimate photograph or film. In addition, the

11.4 Evidence shows that intimate image abuse is a widespread harm, which is increasing in its prevalence online. In 2024, over 22,000 reports of intimate image abuse were made to the Revenge Porn Helpline.[291] The prevalence is particularly high among young adults, with 69% of reports to the Revenge Porn Helpline made by adults aged 18 to 34.[292] The number of intimate image abuse offences recorded by the UK police increased by 40% between 2020 and 2021[293] and total reports to the Revenge Porn Helpline increased by 106% between 2022 and 2023.[294]

11.5 'Deepfake' intimate image abuse is growing particularly quickly, with more deepfake intimate image abuse being posted online in 2023 than in every previous year combined.[295] Ofcom research indicates that 20% of young people have encountered a sexual deepfake.[296]

11.6 Perpetrators of intimate image abuse may be motivated by a variety of reasons, including as part of a pattern of coercive and controlling behaviour against the victim or survivor; to humiliate, shame, and distress the survivor or victim; as part of a 'collector culture' where users collect, share and trade images; or for financial gain ('sextortion').

11.7 Perpetrators can distribute this content on any service with the functionality to post or share images, videos, or files and search services which host or index images, videos, or files. This includes sites dedicated to the non-consensual sharing of intimate images, where perpetrators request, trade and share images.[297]

11.8 We consider the risk of image-based IIA content to be higher on services whose principal purpose is the hosting or dissemination of regulated pornographic content and file-sharing and file-storage services. We set out the specific risks to services in the 'Who this measure applies to' section.

11.9 There is a significant risk of image-based IIA content being re-shared on online services, resulting in the proliferation of this harmful content. For example, the Revenge Porn Helpline indicated that between 2018 and 2022, 35 submissions of image-based IIA content resulted in the identification of 16,937 images being shared online.[298] In many cases, images are shared alongside identifying information about the survivor or victim ('doxing'), which presents a significant risk to the wellbeing and physical safety of these individuals.

11.10 We recognise the significant impacts of intimate image abuse on survivors and victims, across their everyday lives, relationships and professional lives[299] Survivors and victims of intimate image abuse struggle with a range of feelings, including shame, isolation and

Data (Use and Access) Bill as currently drafted introduces new offences of creating or requesting the creation of purported intimate image of an adult without consent. These offences in the Data (Use and Access) Bill would cover doctoring or creating images using artificial intelligence (known as deepfake intimate images).

[291] Revenge Porn Helpline (Papachristou, K.), 2025. Revenge Porn Helpline 2024 Report. [accessed 8 April 2025].

[292] Revenge Porn Helpline (Papachristou, K.), 2024. Revenge Porn Helpline 2023 Report. [accessed 24 April 2025].

[293] Offences recorded by the UK police increased from 3253 (2020) to 4557 (2021). Source: Refuge, 2023. Intimate image abuse – despite increased reports to the police, charging rates remain low. [accessed 27 March 2025].

[294] Revenge Porn Helpline (Papachristou, K.), 2024. Revenge Porn Helpline 2023 Report. [accessed 24 April 2025].

[295] My Image My Choice, 2024. Deepfake Abuse: Landscape Analysis 2023-24. [accessed 1 May 2025].

[296] 20% of young people aged 16 to 24 had come across a sexual deepfake. Source: Ofcom, 2024. Online Nation: 2024 Report. [accessed 1 May 2025].

[297] Henry, N. and Flynn, A., 2019. Image-Based Sexual Abuse: Online Distribution Channels and Illicit Communities of Support, *Violence Against Women,* 25(16). [accessed 13 May 2025].

[298] Huber, A. and Ward, Z., 2024. Non-consensual intimate image distribution: Nature, removal, and implications for the Online Safety Act, *European Journal of Criminology,* 22(1). [accessed 27 March 2025].

[299] Woods, L. and McGlynn, C., 2022. Pornography platforms, the EU Digital Services Act and Image-based sexual abuse, *Media@LSE*, 26 January. [accessed 27 March 2025]; Huber, A., 2023. 'A shadow of me old self': The impact of image-based sexual abuse in a digital society, *International Review of Victimology,* 29(2). [accessed 16 May 2025].

humiliation.[300] The Revenge Porn Helpline reported that 60% of its clients are referred to mental health support services due to the impact of their experiences.[301] Refuge found that intimate image abuse in the context of domestic abuse had resulted in survivors and victims' allowing the perpetrator to have contact with their children or resuming their relationship with the perpetrator.[302] The Illegal Harms Register of Risks (Illegal Harms Register) sets out these impacts on survivors and victims, and notes a reduction in their willingness to particate online.

11.11 Survivors and victims of intimate image abuse experience a constant fear that images will not be removed or might resurface online.[303] Survivors and victims face challenges in reporting intimate image abuse, such as not knowing exactly where the images have been shared or which account shared them.[304] In response, they can develop hypervigilance[305] and many report feeling "extremely fearful" for their safety.[306]

11.12 It is important to note that intimate image abuse is predominately perpetrated against women.[307] The Revenge Porn Helpline identified that women were the survivors and victims in 71% of cases where images were shared without consent.[308] Their data also revealed that women had approximately 28 times more images shared than men.[309]

Our proposal

11.13 To address this harm, we are proposing the following measure to be included in the Illegal Content Codes for certain user-to-user and search service providers:

- Providers should ensure that, where technically feasible, hash matching technology is used to detect image-based IIA content. This involves analysing images communicated publicly on the service and comparing a digital fingerprint (or 'hash') of that content to digital fingerprints (or 'hashes') of previously identified image-based IIA content.

11.14 Refer to Annex 7 and 8 for the full details of our proposed Illegal Content Code measures.

How the measure relates to our Codes of Practice

11.15 Our proposed measure will work in tandem with the existing moderation measures (specifically, user-to-user measures ICU C1 and ICU C2 and search measure ICS C1) in the Illegal Content User-to-User Codes and the Illegal Content Search Codes.

11.16 Following the detection of image-based IIA content, a service provider should treat this as reason to suspect that the content may be illegal content; should review and assess the content; and, if identified as illegal content, take appropriate action on such content in

[300] Law Commission, 2022. Intimate image abuse: a final report. [accessed 27 March 2025].
[301] Revenge Porn Helpline (Papachristou, K.), 2024. Revenge Porn Helpline 2023 Report. [accessed 27 March 2025].
[302] Refuge, 2020. The Naked Threat. [accessed 27 March 2025].
[303] McGlynn, C., Johnson, K., Rackley, E., Henry, N., Gavey, N., Flynn, A. and Powell, A., 2021. 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse, *Social & Legal Studies*, 30(4). [accessed 27 March 2025].
[304] Revenge Porn Helpline (Papachristou, K.), 2025. Revenge Porn Helpline 2024 Report. [accessed 8 April 2025].
[305] McGlynn, C., Johnson, K., Rackley, E., Henry, N., Gavey, N., Flynn, A. and Powell, A., 2021. 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse, *Social & Legal Studies*, 30(4). [accessed 27 March 2025].
[306] Henry, N., Flynn, A. and Powell, A., 2019. Responding to 'revenge pornography': Prevalence, nature and impacts. [accessed 27 March 2025].
[307] Law Commission, 2022. Intimate image abuse: a final report. [accessed 27 March 2025].
[308] Revenge Porn Helpline (Papachristou, K.), 2024. Revenge Porn Helpline 2023 Report. [accessed 27 March 2025].
[309] Revenge Porn Helpline (Papachristou, K.), 2024. Revenge Porn Helpline 2023 Report. [accessed 27 March 2025].

accordance with our moderation measures set out in the Illegal Content User-to-User Codes and Search Codes.[310]

11.17 Our proposed measure complements existing automated content and search moderation measures that tackle child sexual abuse material (CSAM) in the Illegal Content User-to-User and Search Codes, which we set out in our December 2024 statement.

11.18 The complementary measures for User-to-User services are:

- ICU C9 – recommends the use of hash matching technology for certain user-to-user service providers to detect and remove image-based CSAM.[311]
- ICU C10 – recommends the use of URL detection technology for certain user-to-user service providers to detect CSAM URLs.

11.19 The Illegal Content Search Codes also includes a measure (ICS C7) recommending the use of URL detection technology for providers of general search services to remove listed CSAM URLs from search results. We are not recommending the use of URL detection for intimate image abuse because we do not have sufficient evidence to demonstrate its effectiveness in detecting this type of harmful content online. Rather, at this stage, we are recommending that certain search service providers use hash matching technology to detect intimate image abuse. We detail our reasoning in the 'Who this measure applies to' section.

11.20 In addition to these existing measures, we are proposing a measure recommending that certain user-to-user service providers use hash matching technology to detect terrorist content and a principles-based measure for certain service providers to deploy proactive technology to detect certain illegal and harmful content. We detail these proposals in the 'What hash matching is, and how our proposal would work' section below and in Annex 14.

11.21 In developing this consultation, we have taken account what we have learnt about the use of hash matching for CSAM and, where it is appropriate and feasible, we have designed the measures to be consistent with Measure ICU C9 (hash matching for image-based CSAM). However, there are some differences, reflecting both the distinctions between CSAM and intimate image abuse and the context in which we are proposing these measures. We outline the recommendations specific to the measure on hash matching for intimate image abuse in the 'How our proposal would work' section.

11.22 Taken as a whole, our automated content and search moderation measures will work together in building and improving the hash matching ecosystem and protecting individuals from a range of illegal content online.

Proactive technology

11.23 Hash matching is a form of proactive technology.[312] As noted in Chapter 3, there are certain limitations on our ability to recommend the use of proactive technology in the Codes.

[310] For a user-to-user service, appropriate moderation action is to take down illegal content, unless it is currently not technically feasible for them to achieve this outcome. For a search service appropriate moderation is to ensure that illegal search content does not appear in search results or is given a lower priority in the overall ranking of search results. Refer to Volume 2, Chapter 2 and Volume 2, Chapter 3 of our December 2024 Statement on Protecting People from Illegal Harms Online (December 2024 Statement) for further information on the content and search moderation measures.
[311] We are proposing changes to who measure ICU C9 applies to. We detail these proposals in Chapter 13.
[312] As defined in s 231 of the Act.

11.24 We have carefully considered the use of proactive technology recommended in this proposed measure.[313] In doing so we have considered accuracy, effectiveness and risk of bias.[314] For example, we recommend providers configure their technology to strike an appropriate balance between precision and recall, while having regard to various specific factors (detailed in 'Explanation of the measure' and in Annex 6. We also consider the accuracy, effectiveness and lack of bias of perceptual hash matching technology in the 'Rights assessment' section.

What hash matching is, and how our proposal would work

Hash matching technology

11.25 We are proposing that service providers in scope of the measure use perceptual hash matching to detect image-based IIA content. For more information on hash matching technology, including perceptual hash matching, refer to Chapter 8 in this consultation.

11.26 We consider that the use of perceptual hash matching will improve the likelihood of detecting image-based IIA content, as it will be more likely to detect similar but not identical matches to known image-based IIA content. This will likely result in the detection of a larger amount of potentially illegal content and, by extension, the appropriate moderation of such content. This will reduce the harm to survivors and victims and the risk of inadvertent user exposure.

11.27 We expect the use of perceptual hash matching will reduce the spread of this type of harmful content online. However, the use of perceptual hash matching is likely to result in content which is not image-based IIA being erroneously identified. To address this, we have designed the measure to include safeguards which mitigate interferences with human rights, as set out in more detail in the 'Rights Assessment' section below.

Explanation of the measure

11.28 We have designed the proposed measure to materially reduce the non-consensual sharing of intimate images and, in doing so, reduce the associated risks of this harm to individuals. This measure targets image-based IIA content by providing service providers with an effective means of detecting this content which will trigger existing illegal content moderation measures and ensure appropriate moderation action is taken.

11.29 Our proposed measure sets out several recommendations for service providers to successfully detect image-based IIA content. The measure recommends that:

- Service providers in scope of this measure should use, where technically feasible, perceptual hash matching technology to detect image-based IIA content.

[313] In this consultation, we are proposing to introduce a measure into our Illegal Content Codes of Practice for user-to-user services which recommends that providers exclude content indicated potentially to be priority illegal content from recommender feeds (see Chapter 14). Under this measure, service providers should use 'relevant available information' to make this assessment. We define 'relevant available information' to include 'indicators generated by technology used on the service'. Where service providers detect content using hash matching technology in line with our proposals in this chapter, this is an 'indicator generated by technology used on the service' and should therefore be treated as 'relevant available information' for the purpose of our proposed recommender systems measure.

[314] In deciding whether to include a proactive technology measure in a code of practice, we must have regard to the degree of accuracy, effectiveness and lack of bias achieved by the technology in question: Schedule 4 to the Act, para 13(6).

11.30 The content in scope of this measure is photographs, videos or visual images.[315] The measure recommends that service providers ensure the hash matching technology is configured to strike an appropriate balance between precision and recall, with service providers having the flexibility to determine when there is a sufficient similarity between the hash in the database and the hash of the piece of content from the service. We are recommending that service providers conduct regular reviews of the hash matching technology to ensure its efficacy.

11.31 Service providers in scope of this measure should source image-based IIA hashes from an appropriate third-party and/or internal database.

- Dependent on the type of service, providers can either source image-based IIA hashes from (1) a third-party hash database; (2) an internal hash database of content the service provider has reviewed and determined to be image-based IIA; or (3) both. We set out these recommendations in more detail in Annex 14 The proposed measure does not prescribe the use of any specific third-party hash database but sets out conditions for the appropriate selection of a hash database.
- We understand that current third-party intimate image abuse database providers do not verify the content submitted for inclusion in their databases, due to the context-specific nature of intimate image abuse and to protect the privacy of survivors and victims. For this reason, hashes do not have to be sourced from a person with expertise in the identification of image-based IIA content, although in practice the operators of the current third-party database do have this expertise.

11.32 Service providers should treat a match by the hash matching technology as reason to suspect the content is intimate image abuse.

i) When there is reason to suspect that content is intimate image abuse (and is therefore illegal), service providers should follow the existing content and search moderation measures and take appropriate moderation action.[316]

ii) Specifically, service providers should review and assess suspected illegal content to confirm whether the detected content is intimate image abuse. This review process may include automated technologies; however, we are recommending that providers ensure that human moderators review and assess an appropriate proportion of detected content. We are also recommending that providers have regard to various factors in determining this proportion (discussed further in Annex 14). If the content is determined to be illegal, the service provider should swiftly take appropriate action to moderate the content.

11.33 We detail the specifics of these recommendations for service providers in Annex 14.

Effectiveness at addressing risks and benefits

11.34 Many large service providers and some smaller service providers already use perceptual hash matching technology to tackle intimate image abuse on their services.

[315] A proactive technology measure may not recommend the use of technology which operates (or may operate) by analysing user-generated content communicated privately, or metadata relating to such content: See paragraph 13(4) of Schedule 4 of the Act. We have issued Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act.

[316] Illegal Content Search Codes, ICS C1; Illegal Content User-to-User Codes, ICU C1 and ICU C2.

11.35 Evidence suggests that hash matching technology is an effective means to detect image-based IIA content, and, by extension, facilitate the removal of this content.[317] For example, StopNCII.org, a third-party hash matching database, indicated that in three years over a million hashes of images have been created.[318] Several of the service providers that partner with StopNCII.org to implement hash matching for intimate image abuse, including search service providers, have told us that hash matching is an effective tool to reduce image-based IIA content.[319] One service provider reported that it took action on nearly 270,000 matched images following a pilot implementation of hash matching.[320]

11.36 We also have evidence that suggests survivors and victims of intimate image abuse are increasingly submitting images to third-party organisations to be hashed and matched against. For example, StopNCII.org experienced a 130% increase between 2023 and 2024 in the number of hashes created by survivors and victims.[321]

11.37 The measure we are proposing would materially increase the number of service providers implementing hash matching for image-based IIA content. Given its efficacy at detecting image-based IIA content, we consider that wider use of hash matching would lead to a material reduction in the amount of such content that is non-consensually uploaded, shared, and encountered online.

11.38 We consider that such a reduction would result in significant benefits due to the severe harms to survivors and victims posed by this content. It would likely reduce the threat that images might resurface online and reduce the burden on survivors and victims to find and report their images. Finally, we expect there would be a reduction in the number of people inadvertently viewing image-based IIA content online.

11.39 We expect that these benefits will increase over time as intimate image abuse is a growing harm, with deepfake intimate image abuse growing particularly quickly.[322] Therefore, the measure will have long-term effectiveness and benefits.

Impacts and Costs

11.40 We expect service providers may incur one-off set-up costs, as well as ongoing maintenance and operating costs. The main costs are likely to consist of the following:

- one-off costs associated with setting up a hash matching system to detect image-based IIA content;
- cost of maintaining the hash matching system;
- cost of software, hardware and data; and
- cost of reviewing matches, moderating content and dealing with appeals.

11.41 To take a conservative approach, we assume all service providers incur the costs of implementing a hash matching system for the first time.

[317] Domestic Abuse Commissioner's Office response to the 2023 Illegal Harms Consultation, p. 5; Victims Commissioner for England and Wales response to the 2023 Illegal Harms Consultation, p. 6; and Refuge response to the 2023 Illegal Harms Consultation, p. 13.
[318] SWGfL, 2024. StopNCII.org Being Used to Protect Over 1,000,000 Images Online. [accessed 2 April]
[319] Ofcom/Aylo meeting, 17 October 2024; Ofcom/Meta meeting, 30 October 2024; Ofcom/[✂] meeting, [✂].
[320] Microsoft (Gregoire, C.), 2024. An update on our approach to tackling intimate image abuse. [accessed 10 April 2025].
[321] Revenge Porn Helpline (Papachristou, K.), 2025. Revenge Porn Helpline 2024 Report. [accessed 8 April 2025].
[322] My Image My Choice, 2024. Deepfake Abuse: Landscape Analysis 2023-24. [accessed 20 May 2025].

11.42 Our proposed measure recommends that all providers of services whose principal purpose is the hosting or dissemination of regulated pornographic content or file-sharing and file-storage services use a third-party hash database. Therefore, we have considered the costs that are likely to be associated with using StopNCII.org (run by South West Grid for Learning), as we understand this is the only third-party database for image-based IIA hashes currently available to service providers. All other service providers in-scope of the measure have the option to use an internal hash database and therefore may choose to incur the relevant costs associated with this.

11.43 The overall costs to providers from implementing our proposed measure is likely to depend on a number of factors such as number of users, technical complexity, and risk for intimate image abuse.

One-off costs

11.44 The one-off costs associated with setting up a hash matching system to detect image-based IIA content will primarily consist of labour costs. It will require input from software engineers, supported by other professionals such as analysts, product managers, and lawyers.

11.45 Depending on the type of service, the service provider can either choose to build an internal hash matching system or use a third-party application programming interface (API) to access the relevant hash matching functionalities. The work and length of time associated with setting up a hash matching system to detect image-based IIA content is likely to vary across service providers based on the setup chosen, as well as the size and technical complexity of services.

11.46 From previous work on our measure recommending hash matching for CSAM (ICU C9), we estimate that it could take two to 18 months of full-time work to set up a hash matching system. We expect the upper bound (18 months) to reflect the time it might take for a large service provider to build an internal hash matching system, while the lower bound (two months) may be more reflective of a smaller service provider integrating hash matching via a third-party API (such as Hasher-Matcher-Actioner[323]).[324] We estimate that equivalent time is likely to be needed from other staff such as product managers, analysts and lawyers to support in developing and establishing a relevant policy. Based on this, we estimate that the one-off set up costs could range from £19,000 to £350,000, depending on the size and complexity of a service.[325] We expect smaller service providers would incur costs towards the lower end of this range.

Ongoing maintenance costs

[323] Meta (Clegg, N.), 2022. Meta Launches New Content Moderation Tool as It Takes Chair of Counter-Terrorism NGO. [accessed 27 May 2025].

[324] However, we are aware that service providers with a hash matching system already in place are likely to require far less work and time to integrate StopNCII.org's hash database into their existing systems. For example, OnlyFans shared that it took approximately 80 hours of tech and engineering time to integrate StopNCII's database. Another stakeholder, [✂], suggested that it took them over 200 hours (more than five weeks) of engineering work to enable a hash-based solution, as well as work by other professionals to develop the policy. Source: UK Parliament, 2025. Tackling non-consensual intimate image abuse. [accessed 7 March 2025]; Ofcom/[✂] correspondence, [✂].

[325] The lower bound assumes that service providers pay median wages and implement the system in around two months, while the upper bound assumes service providers pay double the median wage and take around 18 months to implement the system. We also apply a 21% uplift to gross wages to account for non-wage labour costs (see Annex 15 'Further detail on economic assumptions and analysis' for more details).

11.47 The ongoing maintenance costs will primarily consist of labour costs associated with activities such as applying updates, reviewing the hash matching technology's performance and configuring the technology (at least every six months), ingesting new hashes from the third-party database, and integrating with new service functionalities. This will require input from software engineers as well as other professional occupation staff.

11.48 In line with our standard assumption (see Annex 15 'Further detail on economic assumptions and analysis' for more details), we assume that the annual maintenance costs would be 25% of the one-off costs of setting up a hash matching system for intimate image abuse. Based on this, we estimate that the annual cost of maintaining a hash matching system for intimate image abuse to range from £5,000 to £88,000.

Ongoing software, hardware and data costs

11.49 The ongoing software, hardware and data costs to service providers will mainly consist of the costs associated with accessing hash databases and API solutions from third-party organisations where relevant. As these costs generally tend to scale with a service's user base and/or revenues, we expect large service providers to pay more than smaller service providers. In 2025, StopNCII.org began charging for access to its hash database. The amount a service provider has to pay is determined based on their annual revenue and can start at around £1,000 per year. We are aware that StopNCII.org are willing to work with service providers with less ability to pay, to provide lower cost options.[326] There are free open-source third-party API solutions available to service providers, although these solutions require access to computing power to check for content on services. We assume the annual cost of software, hardware and data could start at around £2,000 for a small service provider.

Ongoing content and search moderation and appeal costs

11.50 The ongoing content and search moderation and appeal costs will mainly consist of labour costs associated with reviewing an appropriate proportion of content that is matched by the hash matching technology to ensure the efficacy of the technology, moderating the detected content and dealing with appeals. These costs are likely to scale with the amount of content that is matched by the hash matching technology. Therefore, we expect that service providers with a small user base and lower risk for intimate image abuse are unlikely to require a full-time content moderator, relative to a service provider with a high reach and/or high risk for intimate image abuse.

11.51 As an example, we estimate that a service that receives around 80 matches per year for review could spend around £400 to £1,000 on human moderation per year, if undertaking a review of every match (although providers have the flexibility to review an appropriate proportion of content that is matched).[327] Some services may have a much lower or higher number of matches. We generally expect that the ongoing cost of content and search

[326] South West Grid for Learning (Wright, D), 2025. StopNCII Scalability, Capacity, and Suitability for Mandated Hashing Technology Implementation. [accessed 10 April 2025].

[327] Based on our stakeholder engagement, we understand that it is difficult for service providers to specifically identify the time moderators may spend reviewing content detected as image-based IIA. As a proxy, we instead use global data on the number of content enforcements made and the number of moderators employed by Snapchat, to estimate a ratio for the number of moderators per report (0.0001). Applying this ratio, we expect a service with 80 matches for review to employ the full-time equivalent of around 0.01 moderators. We then apply our assumptions about the salary of a content moderator and the uplift for staff costs, as set out in Annex 15 'Further detail on economic assumptions and analysis'.

moderation will correlate with the risk and amount of intimate image abuse on a service, and therefore would be in line with the benefits of our proposed measure.

11.52 Among those that have implemented hash matching using StopNCII.org's database the number of matches identified can vary across service providers. One stakeholder Aylo indicated that the number of matches over a period of three months was very small. Another stakeholder [✂] shared that it had lots of matches and a lot of false positives during the initial integration of the StopNCII.org database.[328] Although the costs of moderating this detected content could be high during initial integration, we expect these costs to fall over time as service providers are better able to control (to an extent) for the volume of erroneously detected content when configuring their technology.

11.53 Once service providers have reason to suspect content is image-based IIA, they will need to moderate the content in accordance with our existing content and search moderation measures. Service providers will also need to respond to any appeals from complainants whose content had been incorrectly moderated because of the hash matching technology.[329] We expect the upfront costs associated with these processes to already be covered by existing code measures (content moderation, search moderation and reporting and complaints) which apply to all service providers. However, we acknowledge that our proposed measure may potentially result in more content being identified for moderation and more appeals.

11.54 As mentioned above, to take a conservative approach, we have assumed that service providers will incur the costs of implementing a hash matching system for the first time. However, we note that our proposed measure is recommended for user-to-user service providers that are already likely to be within scope, or may become within scope,[330] of an existing measure recommending the use of hash matching to detect and remove CSAM (ICU C9);[331] and some providers of large general search services (such as Google Search and Microsoft Bing) that may already use hash matching to detect illegal content on their services.[332] Therefore, the costs above are likely to be an overestimation of the costs that may be incurred by some service providers in scope of our proposed measure, as we expect that:

- Service providers that have already implemented hash matching to detect CSAM would incur significantly lower additional costs if also implementing hash matching for image-based IIA content.
- Service providers that may need to implement hash matching to detect CSAM going forward may benefit from cost synergies if also implementing hash matching for image-based IIA content.
- Large general search services that already use hash matching to detect illegal content on their services are likely to have already incurred most of the relevant costs of our proposed measure.

[328] Ofcom/Aylo meeting, 17 October 2024; and Ofcom/[✂] meeting, [✂].

[329] Refer to Volume 2, Chapter 6 of our December 2024 Statement for further information on the reporting and complaints measures.

[330] We are proposing changes to who measure ICU C9 applies to. We detail these proposals in Chapter 12.

[331] We expect there to be a strong overlap between services that are at risk of CSAM and services that are at risk of intimate image abuse.

[332] Google (Jasper, S.), 2022. How we detect, remove and report child sexual abuse material. [accessed 10 April 2025]; Microsoft (Gregoire, C.), 2024. An update on our approach to tackling intimate image abuse. [accessed 7 April 2025].

Rights assessment

11.55 There is a substantial public interest in introducing a measure that aims to reduce the online prevalence and dissemination of image-based IIA content. That public interest relates to the prevention of crime, the protection of health and morals, and the protection of the rights of others. The UK Parliament has now enshrined the sharing of, or threatening to share, intimate images without consent as offences under the Act, making intimate image abuse priority illegal content. The proposed measure is designed to reduce and deter the commission of these offences and thereby reduce the viewing of illegal content of this type. The measure is also designed to protect the right of users to privacy and to respect for a private and family life (Article 8 of the European Convention on Human Rights (ECHR)), which is impacted by the online prevalence and distribution of intimate image abuse.

11.56 As part of our rights assessment, we must consider how the proposed measure may interfere with human rights, and whether any such interference is proportionate. We recognise that content and search moderation functions can have particular impacts on individuals' and entities' rights to privacy under Article 8 of the ECHR and to freedom of expression under Article 10 of the ECHR. Any interference with the right to privacy or the right to freedom of expression must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest. In either case, to be necessary, the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.

11.57 In considering whether impacts on these rights are proportionate for the purposes of this measure, our starting point is to recognise that the detection and moderation of image-based IIA content acts directly to prevent crime by deterring users from sharing such illegal content. It similarly acts to protect public morals, including by reducing the possibility of users inadvertently encountering image-based IIA content online.

11.58 We recognise that the removal or (in the case of search) downranking or delisting of image-based IIA content also acts directly to protect the privacy rights of survivors and victims of abuse and their personal data.

Freedom of expression

Potential interference with freedom of expression

11.59 If action is taken on content that is not illegal based on an erroneous match for image-based IIA content, there may be an interference with users' right to freedom of expression. The action taken may include banning an account. This measure recommends the use of perceptual hash matching, which (compared with cryptographic hash matching) is more likely to detect image-based IIA content but is also more likely to incorrectly detect content as a match for illegal content.

11.60 Unlike our CSAM hash matching measure (ICU C9), we are not recommending that the database be sourced from one or more persons with expertise in the identification of this content. We note that the hash database ecosystem for intimate image abuse is not as developed as that for CSAM, and the offences differ in nature.

11.61 It should be noted that, as discussed in the 'Source image-based IIA hashes' section of this chapter, current third-party intimate image abuse databases do not verify the content submitted by survivors and victims for inclusion in the database.[333] This means there may be greater risk that content included in the database is not image-based IIA. This is borne out by the fact that, as noted above in the 'Costs' section of this chapter, one stakeholder shared that it had lots of matches and a lot of false positives during the initial integration of the StopNCII.org database.

11.62 We understand there is a risk that unverified hash databases could be used maliciously with the aim of targeting content online for moderation, and this risk may be higher for consensually shared pornographic content.

11.63 As this measure allows flexibility in the configuration of the perceptual hash matching technology, there is a risk that its use could lead to varied impacts on users' freedom of expression.

Safeguards and proportionality

11.64 Perceptual hash matching is appropriate to address the harm of intimate image abuse because, unlike cryptographic hash matching, it allows modifications from an original file to be detected, which should render attempts to evade detection by modifying the file ineffective. We consider that the recommendation for service providers to human review an appropriate proportion of detected content is sufficient to mitigate the risk of incorrectly detected images.

11.65 We also note that the only existing third-party intimate image abuse hash database, StopNCII.org, does not verify whether hashes in the database are created from image-based IIA content. This is to maintain the privacy of survivors and victims who create the hash on their device and share it with the database provider. Consequently, it is possible that the database may contain some images which are not intimate images. We understand, in the rare instances images which are not intimate have been detected, it is a result of images being submitted by survivors and victims as a means of testing the process before submitting their intimate images to be hashed. We are not aware of any evidence of unverified hash databases being used maliciously with the aim of targeting content online for moderation. While we understand the risk, we are not aware that it has materialised on services which use hash matching to tackle intimate image abuse.

11.66 Our proposed measure recommends that a match with a hash from an unverified hash database should be treated by providers as reason to suspect that detected content may be image-based IIA content. When there is reason to suspect content may be image-based IIA, the provider should review, assess and take appropriate content moderation action on the identified content, in line with recommended content moderation measures (ICU C1 and ICU C2 and ICS C1). As part of this process, providers should ensure that human moderators review and assess and appropriate proportion of detected content, having regard to factors set out in Annex 14. We consider the content moderation measures, and our recommendation in relation to human review, will help mitigate the risks arising from lack of verification in the database by ensuring the review and assessment of content suspected to be image-based IIA content.

[333] Currently, the only existing third-party intimate image abuse hash database, StopNCII.org, does not verify whether hashes in the database are created from image-based IIA content. This is to maintain the privacy of survivors and victims who create the hash on their device and share it with the database provider.

11.67 We also consider that we can sufficiently mitigate the risks discussed in the 'Potential interference with freedom of expression' section through certain safeguards. These include the following recommendations:

- The source of hashes (the database) gives service providers reason to suspect that detected content might be image-based IIA content.
- Service providers should consider principles relating to the configuration of the technology to ensure that the hash matching technology strikes an appropriate balance between precision and recall. The balance should take into account the risk of harm from intimate image abuse, the proportion of content which is not intimate image abuse detected as a match by the technology, and the effectiveness of the systems and processes used by the service provider to address this content. These principles aim to help ensure the accuracy and effectiveness of the technology and should therefore limit the risk of detection of content that is not intimate image abuse.
- Hash matching technology should be reviewed to ensure continued precision and recall in the detection of content that is intimate image abuse.
- The database should be protected from security compromises through attacks by bad actors.

11.68 We consider the fact that the measure allows flexibility in the configuration of the perceptual hash matching to be a benefit in terms of enabling the measure to be adopted more widely. While we are aware that the impact on freedom of expression may vary according to the approach taken by each service provider, in practice service providers have reputational and commercial incentives to encourage them to strike a reasonable balance in their implementation of the measure.

11.69 While providers have incentives to minimise the amount of image-based IIA content on their service and to limit the costs of human moderation, incentives also exist to limit the amount of content that is wrongly taken down (or in the case of search, removed from search results or given lower priority) and reduce the costs of handling user complaints, including appeals.

11.70 Interference with users' freedom of expression which arises where providers take action against users (such as banning an account) because the user has been detected as sharing image-based IIA content. We consider that impact in Chapter 16. The safeguards included in this measure to protect users' freedom of expression would limit the risk that action is taken against users on the basis of false positives.

11.71 Overall, we acknowledge that the measure involves interference with users' rights to freedom of expression where content is incorrectly detected as image-based IIA content. However, we consider the interference to be limited having regard to the safeguards in place and proportionate to the measure's aim of reducing the volume of image-based IIA content being shared and viewed online.

Privacy and data protection

Interferences with privacy and data protection

11.72 We have considered the impact on users' right to privacy under Article 8 ECHR more broadly.

11.73 Hash matching will often involve automated processing at scale of images or videos uploaded, generated, or shared by users. This will often involve processing of personal data, either because users or other individuals are identifiable from the content, or because the content is connected to other information (such as an account profile), which renders someone identifiable.

11.74 According to the Information Commissioner's Office (ICO), automated processing of personal data can lead to a number of possible data protection harms, such as loss of control of personal data, 'invisible processing' or unwarranted surveillance.[334]

11.75 The proposed measure involves the use of proactive technology to detect image-based IIA content, and the use of human moderators to review an appropriate proportion of detected content. This measure would only apply to content that is communicated publicly, which is consistent with the Act's constraints on proactive technology. Because of this, the user sharing the content should have a reduced expectation of privacy in connection with the content. That said, people depicted in the content are likely to have a reasonable expectation of privacy (for the purposes of Article 8 ECHR) in relation to content that is considered to be communicated publicly for the purposes of the Act, because it may have been shared without their consent. To the extent that the measure involves an interference with that person's right to privacy, we discuss why we consider this to be proportionate below.

11.76 The proposed measure recommends the automated scanning of relevant content to detect hash matches for image-based IIA, which could result in the processing of personal data of individuals.

11.77 This measure also involves human review of an appropriate proportion of content by service providers to ensure the efficacy of the hash matching technology.

Privacy and data protection safeguards

11.78 Intimate image abuse is a significant privacy violation, and this measure will contribute to protecting the privacy rights of survivors and victims.

11.79 This measure would only apply to content that is communicated publicly, which is consistent with the Act's constraints on proactive technology.

11.80 In the context of this measure, the impact on the privacy is mitigated by the fact that the image has been submitted in order for it to be processed in this way by providers, for the aim of preventing or reducing its public circulation.

11.81 Where service providers ensure that the automated processing involved in hash matching is carried out in accordance with data protection law, that processing should accordingly have a minimal impact on users' privacy.

11.82 We consider that viewing of the image by human moderators when there is a match does not have undue impact on the privacy of the person who *posted* the content, as at the point the hash is matched, the image will have been communicated publicly by that person. Human review may represent an interference with the privacy rights of the survivors and victim depicted in the image. However, where that person has submitted the image to the database, they have essentially agreed to consideration of its content for the purpose of

[334] ICO response to November 2023 Consultation on Protecting People from Illegal Harms Online, p.12. See also: ICO, 2022. Overview of Data Protection Harms and the ICO's Taxonomy. [accessed 18 October 2024].

reducing or preventing public distribution. In addition, that review forms an important part of ensuring that the overall measure – which aims to reduce the serious impact on the privacy of survivors and victims'– is proportionate and effective. We therefore consider that the intrusion into survivors and victims' privacy rights is necessary, and that no less intrusive approach would be a suitable alternative.

11.83 In designing measures, we have borne in mind the importance of ensuring service providers comply with both online safety and data protection rules. Service providers will need to ensure that the automated processing involved in using hash matching technologies, and all other associated processing (such as review of detected content by human moderators) is carried out in accordance with applicable data protection law, including the UK GDPR.

11.84 The UK GDPR places a specific restriction on making decisions based solely on automated processing of personal data, where the decision has legal or similarly significant effects for the relevant individual. The ICO has set out that decisions to take down content can (in some cases) have such effects.[335] So-called 'automated decision-making' is only permitted where service providers have implemented certain safeguards for the data subject's rights, freedoms, and legitimate interests in accordance with the UK GDPR.[336] Service providers should have regard to the ICO's guidance on automated decision-making and profiling when implementing this measure.[337]

11.85 We are satisfied that the processing required by the measure can be carried out in accordance with data protection law. The ICO's guidance on content moderation and data protection[338] advises service providers to carry out a data protection impact assessment to assess and mitigate data processing risks.

11.86 We have also identified other existing measures as safeguards for users' privacy (including the protection of personal data). These include ensuring prospective complainants can appeal if they believe a provider has taken incorrect action on their content due to being wrongly identified as intimate image abuse.

11.87 Overall, while we acknowledge the measure may involve some interference with users' right to freedom of expression where incorrect detection of content results in action being taken by providers and right to privacy (as outlined above), we consider that the safeguards set out above sufficiently mitigate any such impact. We consider that any such interferences are proportionate to the measure's aim of reducing the prevalence and dissemination of intimate image abuse.

Which providers should implement this measure

11.88 Our Risk Assessment Guidance, recognises that not all services will have the same risk of intimate image abuse and sets out the factors which we consider indicate that a service should be assessed as being at low, medium or high risk for intimate image abuse.

11.89 We are proposing to target the measure based on a service's risk and number of users. We are therefore provisionally recommending the following service providers use perceptual hash matching for intimate image abuse:

[335] ICO, 2024. Content moderation and data protection. [accessed 24 April 2025].

[336] See Article 22(3) of the UK GDPR and section 14 of the Data Protection Act 2018.

[337] ICO, n.d. Automated decision-making and profiling. [accessed 24 April 2025].

[338] ICO, 2024. Content moderation and data protection. [accessed 24 April 2025].

i. Providers of user-to-user services which are **high risk** of intimate image abuse and either:

   - whose **principal purpose is the hosting and dissemination of regulated pornographic content**; or
   - are **file-sharing and file-storage** services; or
   - have **more than 700,000** monthly active UK users.

ii. Providers of **large** user-to-user services that are **medium risk** for intimate image abuse.

iii. Providers of **large** general **search** services.

11.90 These service providers will be expected to use a third-party or internal hash database for this measure (depending on which category they fall into).

11.91 We are recommending that the following service providers will need to use, at minimum, a third-party hash matching database. These service providers can use an internal hash database, in addition to the third-party database, if desired.

- Providers of user-to-user services which are **high risk** of intimate image abuse and either:
  - whose **principal purpose is the hosting and dissemination of regulated pornographic content**; or
  - are **file-sharing and file-storage** services.

Assessment for application to specific service providers

Providers of user-to-user services which are at high risk of intimate image abuse and whose principal purpose is the hosting or dissemination of regulated pornographic content

11.92 We provisionally consider that it would be proportionate to bring providers of services whose principal purpose is the hosting or dissemination of regulated pornographic content in scope of the measure, regardless of monthly UK user base.

11.93 There is a high risk of intimate image abuse on user-to-user services whose principal purpose is the hosting or dissemination of regulated pornographic content, unless the service provider uses hash matching or an equivalent safety measure to tackle intimate image abuse.

11.94 The Illegal Harms Register highlights that significant amounts of intimate image abuse are hosted on pornography services. These services hosted 25% of all content reported to the Revenge Porn Helpline in 2024, more than any other service type.[339]

11.95 Pornography sites are high risk for intimate image abuse regardless of size. Evidence suggests that smaller pornography sites may ‘pull’ content from larger pornography sites, which enables image-based IIA content to proliferate quickly.[340] Research from the Australian eSafety Commissioner found that less visible user-generated pornography sites

[339] Revenge Porn Helpline (Papachristou, K.), 2024. Revenge Porn Helpline 2024 Report. [accessed 27 March 2025].

[340] Huber, A. and Ward, Z., 2024. Non-consensual intimate image distribution: Nature, removal, and implications for the Online Safety Act, *European Journal of Criminology,* 22(1). [accessed 7 April 2025].

host thousands of non-consensual images. A search for 'revenge' on one such site produced over 20,000 results.[341]

11.96 Services whose principal purpose is the hosting or dissemination of regulated pornographic content will likely have large volumes of nudity and sexually explicit content and are unlikely to have moderation systems to detect and remove nudity and sexually explicit content. Without additional context such as a user report it can be challenging to determine whether such content was shared without consent. However, dependence on user reports is likely to be unreliable, because survivors and victims are unlikely to be aware of this content. These challenges could impact the efficacy of an internal database. We therefore consider it appropriate to recommend the use of a third-party database for providers of services whose principal purpose is the hosting or dissemination of regulated pornographic content.

Providers of user-to-user services which are at high risk of intimate image abuse and are file-sharing and file-storage services

11.97 We provisionally consider that it would be proportionate to bring providers of file-sharing and file-storage services in scope of the measure, regardless of monthly user base.

11.98 As outlined in the Illegal Harms Register, there is strong evidence to suggest that file-sharing and file-storage services (including image sharing services) are particularly high risk for intimate image abuse. For example, the Revenge Porn Helpline identified that file-sharing services are a primary location for the collection of, and sharing of, intimate images. In one case, 940 images were found in a collection on a file-sharing site and were able to be shared as a single file.[342]

11.99 The risk of harm from illegal content on file-sharing and file-storage services does not necessarily scale with the number of users, as many file-sharing and file-storage services allow users to share large numbers of images via a single link. Perpetrators of intimate image abuse use file-sharing and file-storage services to create links to content and subsequently share these links on other user-to-user services.[343]

11.100 Furthermore, only a small number of file-sharing and file-storage services reach more than 700,000 monthly UK users and our analysis shows that user numbers can fluctuate substantially for smaller file-sharing and file-storage services. If we set the same user number threshold as for other service types, this would likely mean the proposed measure would not apply to many file-sharing and file-storage services.

11.101 The prevalence of image-based IIA content on file-sharing and file-storage services is likely to be under-reported because content that is shared publicly (such as files) on the services can be challenging for individuals to locate online. This means survivors and victims may not be aware of the content being shared on these types of services and, therefore, are unable

[341] Office of the eSafety Commissioner, 2017. Image-Based Abuse. National Survey: Summary Report. [accessed 13 May 2025].

[342] Huber, A. and Ward, Z., 2024. Non-consensual intimate image distribution: Nature, removal, and implications for the Online Safety Act, *European Journal of Criminology,* 22(1). [accessed 7 April 2025].

[343] Recent analysis of case data from the Revenge Porn Helpline identified file-sharing services as a location for the collection of and sharing of intimate images, where large collections of images can be made and shared with a single link. Further, file-sharing is linked with discussion forums and chatrooms, where chat threads will use file-sharing links to share large amounts of IIA content in forums and threads. Source: Revenge Porn Helpline (Huber, A. and Ward, Z.), 2024. Non-Consensual Intimate Image Distribution: Nature, Removal, and Implications for the Online Safety Act. [accessed 5 June 2025].

to report it to these service providers.[344] We consider the risk of under-reporting may impact the efficacy of an internal hash database. For this reason, we consider it appropriate to recommend the use of a third-party database for providers of file-sharing and file-storage services.

Providers of user-to-user services which have more than 700,000 monthly active UK users and are high risk for intimate image abuse.

11.102 Given the prevalence and impact of intimate image abuse and the fact that hash matching is an effective tool for combatting this harm, we provisionally consider that it would be proportionate include all high-risk services with more than 700,000 monthly users in the scope of this measure.

11.103 In principle, we consider that it may be proportionate to recommend high risk services with fewer than 700,000 monthly active UK users use hash matching for intimate image abuse. This assessment is based on the likely risk of intimate image abuse being on smaller services who are at high risk of intimate image abuse due to their functionalities or service type. However, while the evidence on the risk posed by pornography and file sharing services is clear, the evidence on intimate image abuse on other service types is less developed. In addition, this would significantly increase the number of service providers required to use a third-party hash-database and, at this time, we do not believe the hash-matching ecosystem is developed enough to onboard this many services in a timely manner to facilitate compliance. We will continue monitoring the hash-matching ecosystem and consider extending the measure to other services at a later date.

Providers of large user-to-user services which are at medium risk of intimate image abuse

11.104 We provisionally consider that it would be proportionate to recommend that all providers of large services which pose a medium risk of intimate image abuse should hash match for such content. Given their size, very significant numbers of people could see intimate image abuse that circulates on large services. To protect victims and survivors' privacy, it is therefore important that this measure applies to large medium risk services.

Providers of large general search services

11.105 We provisionally consider it would be appropriate to bring providers of large general search services in scope of this measure.

11.106 We consider that there is potential for widespread circulation of image-based IIA content via image search results on large general search services. Image search functionalities can display image-based IIA content and allow perpetrators to search for this type of content and identify survivors and victims. The benefits of applying this measure to large general search services are therefore likely to be significant.

11.107 Several large general search service providers already use hash matching technology to detect image-based IIA content on their services.[345] We expect that hash matching technology will support search service providers in automating their detection of image-

[344] We are aware of a number of reports of intimate images being shared on file-sharing or file-storage services and it does not appear as though the perpetrators intend for the survivor or victim to know that the image has been shared. As set out in the Illegal Harms Register, this is sometimes referred to as 'collector culture' and is in contrast to other forms of intimate image abuse where the perpetrator shares the images to directly threaten, coerce or humiliate the survivor or victim.

[345] Microsoft (Gregoire, C.), 2024. An update on our approach to tackling intimate image abuse. [accessed 7 April 2025]; Ofcom/Google] meeting, 12 December 2024.

based IIA given the large quantity of content (including pornographic content) that appears in image search on their services. For example, Microsoft Bing reported it had identified and acted on 268,899 images which were being returned in image search results in Bing following a pilot implementation of StopNCII.org’s intimate image abuse database.[346]

11.108 For these reasons, we consider the use of hash matching on large general search services is necessary to facilitate the detection and moderation of image-based IIA content on these services.

Provisional Conclusion

11.109 In this chapter, we have set out evidence suggesting that wider use of hash matching technology would reduce the harm from intimate image abuse. As intimate image abuse is a growing harm, we expect that the benefits of this measure will increase over time.

11.110 Our analysis suggests that the use of hash matching would impose costs anywhere between the low tens and hundreds of thousands of pounds on service providers in scope of this measure, and that costs would be significantly lower for service providers already utilising hash matching technology. Given the severity and scale of intimate image abuse as a harm, we provisionally consider that these costs are proportionate.

11.111 While the proposed measure could have an impact on freedom of expression rights, we have taken several steps to mitigate this impact and consider that any residual impact would be warranted given the anticipated benefits of the measure in reducing the prevalence and dissemination of image-based IIA content.

[346] Microsoft (Gregoire, C.), 2024. An update on our approach to tackling intimate image abuse. [accessed 20 May 2025].

12. Perceptual hash matching for terrorist content

Summary

Terrorists use and exploit online services in a range of ways. This includes disseminating large volumes of terrorist materials, encouraging others to carry out terrorist acts, and attempting to recruit adults and groom children to join proscribed terrorist organisations. Online services are also used to organise and coordinate terrorist attacks, potentially leading to mass casualties.

Our Illegal Harms Codes already recommend both that service providers make it easy for users to report terrorism content, and that they take such content down swiftly when they become aware of it. However, while user reporting is important, there are limitations to the role it can play. Users are not always able to easily spot terrorism content. We also know that many users, including children, fail to report such content – either because they find it difficult to navigate reporting processes, or do not believe the service provider will take action. Lastly, given the scale of terrorism content uploaded on the largest services, user reporting and human content moderation on their own are not likely to be adequate tools.

To address these limitations, we are proposing that large user-to-user services that are at medium or high risk of terrorism content, and services which are at high risk of terrorism content and have over 700,000 monthly UK users (or are file-storage and file-sharing services), should use hash matching technology to detect known terrorism content, so that it can be removed swiftly. This would reduce the circulation of terrorism content and help combat the harm it causes.

Our proposals

Number	Proposed measure	Who should implement this
ICU13	Providers use perceptual hash matching to detect terrorism content so that it can be removed.	Providers of the following user-to-user services that enable regulated user-generated content in the form of photographs, videos or visual images (whether or not combined with written material) to be generated, uploaded or shared: • Large services at medium or high risk of terrorism content. • Other services which are at high risk of terrorism content and: > Have more than 700,000 monthly active UK users; or > Are file-storage and file-sharing services.

Consultation questions

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

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

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

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

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

What risks does terrorism content pose?

12.1 Terrorists use online services to organise, recruit, fundraise and otherwise disseminate terrorism content.347 Terrorists do not rely on a single service but use many services and their associated functionalities. We have found that there is often cross platform sharing of terrorism content from being first posted on small user-to-user services and then linked to on larger, higher reach services and vice versa.[348]

12.2 Large volumes of terrorism content circulate online – a significant portion of this, notably content aimed at propagating terrorist ideology, circulates on public online services. Almost all UK terrorist attack planners from 2012 – 2017 downloaded, shared or consumed content and activity associated with terrorism or extremism online.[349] In addition, research by Tech Against Terrorism identified terrorist content on a total of 187 different tech platforms between November 2020 and January 2023. This resulted in the identification and sending of 22,615 terrorist content URLs in alerts to 95 different tech platforms.[350]

12.3 Terrorism content is illegal and causes significant harm. Our Register of Risks outlines how exposure to it can induce fear, anxiety or panic in a population or subset of a population.[351]Research by the Prison and Probation Service also reported that the internet plays an increasingly prominent role in radicalisation processes, in line with wider society's increasing use of the internet.[352]

12.4 Whilst user reporting is useful, there are limitations to the role it has to play in detecting terrorism content. Users are not always able to easily spot terrorism content, and we know that many users do not report content because they find it difficult to navigate reporting processes or because they do not believe service providers will take action.[353] Where content is directed to terrorist sympathisers or supporters, they are unlikely to report it.

[347] Terrorism content is defined in section 59(8) of the Online Safety Act 2023 (the Act) and means content that amounts to an offence specified in Schedule 5 to the Act.

[348] Ofcom Illegal Harms Statement, Register of Risks, 2024, p.36.

[349] Home Office Interim Code of Practice on Terrorism Content and Activity Online December 2020 [accessed 25 April 2025]

[350]Tech Against Terrorism Patterns of Online Terrorist Exploitation 2023 [accessed 16th April 2025]

[351] Ofcom Illegal Harms Statement, Register of Risks, 2024, p.38.

[352] HM Prison and Probation Service (Kenyon. J, Binder, J, and Baker-Baell, C.), 2022, The Internet and radicalisation pathways: technological advances, relevance of mental health and role of attackers [accessed 16 April 2025]

[353] Ofcom Online Nation Report 2022, p.73 [accessed 16 April 2025]

Due to the volume of content on many services, user reporting and human content moderation alone are not able to detect and remove terrorism content sufficiently at scale.

12.5 In our November 2023 Consultation on Protecting People from Illegal Harms Online (November 2023 Consultation), we asked for evidence on the deployment of hash matching technology to detect terrorism content and its associated accuracy, effectiveness, costs and impact on users' rights.[354] We received responses from industry and civil society organisations confirming that many services already use hash matching to detect terrorist or violent extremist content.

Our proposals

12.6 We are proposing that providers of certain user-to-user services use perceptual hash matching technology to analyse regulated user-generated content in order to assess whether it is terrorism content, for the purpose of complying with their illegal content safety duties under section 10(2) and (3) of the Online Safety Act 2023 (the Act).[355]

12.7 Where content is detected by hash matching, providers should treat this as reason to suspect that the content may be illegal content. Detected content should therefore be reviewed and assessed in accordance with Measure ICU C1 to determine whether it is illegal content or illegal content proxy.[356] If so, it should be swiftly taken down in accordance with Measure ICU C2. Providers' systems and processes should be operated to ensure human moderators review and assess an appropriate proportion of detected content, having regard to certain factors.

12.8 The content that we propose should be analysed by hash matching is regulated user-generated content in the form of photographs, videos or visual images that is communicated publicly by means of the service. For the avoidance of doubt, such content should be analysed whether or not the photographs, videos or visual images are combined with written material. At this stage we are not proposing to include audio content in the scope of the measure (see paragraph 12.13 for further details on our reasons for not including this content).

12.9 We are proposing that the measure should apply to providers of user-to-user services that enable user-generated content in the form of photographs, videos or visual images (whether or not combined with written material) to be generated, uploaded or shared and:

- are large services at medium or high risk of terrorism content; or
- are at high risk of terrorism content and:
  - > have more than 700,000 monthly active UK users; or
  - > are a file-storage and file-sharing service.

12.10 This proposed measure builds on and complements measures which are already in our Illegal Content User-to-user Codes. We have designed this measure to align with the existing measure relating to hash matching for CSAM (Measure ICU C9). For example, the measures recommend the same steps to ensure accuracy, prevention of bias, and security

[354] Ofcom Illegal Harms Consultation, Draft Illegal Harms Codes of Practice, 2023, p.118, paragraphs 14.138 – 14.140.

[355] Regulated user-generated content is defined in section 55(2) of the Act.

[356] Illegal content proxy is content that a provider determines to be in breach of its terms of service, where: a) the provider had reason to suspect that the content may be illegal content; and b) the provider is satisfied that its terms of service prohibit the type of illegal content which it had reason to suspect existed.

of the database. However, the proposed measure for terrorism content recommends that content detected by hash matching should always be reviewed and assessed to determine whether it is illegal content or illegal content proxy. This is because of the importance of context when identifying terrorism content. In paragraph 12.72, we provide further detail on this important safeguard for users' rights.

Explanation of the Measure

12.11 We propose recommending that, where technically feasible, providers should use perceptual hash matching technology to analyse regulated user-generated content in the form of photographs, videos or visual images (whether or not combined with written to assess whether it is likely to be terrorism content. Only content communicated publicly by means of the service should be analysed, in accordance with the constraints on our power to recommend proactive technology measures.[357]

12.12 All relevant content present on the service at the time the technology is implemented should be analysed within a reasonable time. Relevant content generated on, uploaded to, or shared on the service (or that a user seeks to generate, upload or share) after the technology is implemented should be analysed before or as soon as practicable after it can be encountered by UK users.

12.13 Terrorist groups have long produced propaganda and other online content including photographs, videos and visual images, which may be combined with written material, for example in magazines and instructional material. While audio content has also been exploited by terrorist groups for many years,[358] [359] current research indicates that the availability of technical tools to detect audio content across different services varies dramatically. In addition, there are challenges related to the availability of appropriate training data, and the amount of time, high degree of subject matter expertise, and contextual knowledge required to detect and assess audio content.[360] [361] Therefore, at this stage, we are not proposing to include audio content in the scope of this measure. However, we would welcome evidence from stakeholders on the use of hash matching technology to tackle terrorism content in audio form in response to the consultation.

12.14 Our proposed measure would only apply where it is technically feasible for a provider to implement the technology. We do not consider that it would be technically infeasible to implement a measure merely because to do so would require some changes to be made to the design and/or operation of the service. However, the measure does not apply to providers for whom it is not technically feasible to analyse user-generated content[362]

[357] Schedule 4, paragraph 13 to the Act.

[358] Institute for Strategic Dialogue The terrorist radio revival: How the Islamic State's radio station survives on social media January 2024 [accessed 16 April 2025]

[359] Institute for Strategic Dialogue: The 'original sounds' of terrorist leaders: A TikTok feature enables terrorist content to flourish July 2024 [accessed 16 April 2025]

[360] Global Internet Forum to Counter Terrorism: Broadening the GIFCT hash sharing database taxonomy: An assessment and recommended next steps July 2021 Page 19 [accessed 16 April 2025]

[361] Global Internet Forum to Counter Terrorism: Advances in hashing for counter terrorism March 2023 [accessed 16 April 2025]

[362] This is consistent with our approach to application of Measure ICU C9 ('Using hash matching to detect and remove CSAM') – see paragraph 4.69 of our Statement on protecting people from illegal harms online.

12.15 The measure includes several recommendations relating to the design and configuration of the hash matching system designed to ensure that it is (so far as possible) accurate, effective and free of bias.[363] For example, as explained below, we recommend providers configure their technology to strike an appropriate balance between precision and recall, having regard to various specific factors. We also set out recommendations relating to the conditions of the hash database, and review and assessment of detected content.

Use of hash matching technology

12.16 The detection of terrorism content poses significant challenges due to the way terrorist organisations and their supporters constantly seek to evade trust and safety or content moderation operations through the manipulation of user generated content. Research by the Global Network on Extremism and Technology outlines some of the tactics used by terrorist actors or their supporters to maintain a presence on large services, including the use of image obscuration (for example, cropping, filtering or otherwise editing well-known terrorism content), terrorist symbology (such as logos, flags or emblems) and other evasion behaviour.[364]

12.17 As explained in Chapter 8 in the context of harmful content detection, cryptographic and perceptual hash matching are typically most relevant. We propose recommending that service providers use perceptual hash matching technology because, in practice, it is likely to detect a larger amount of terrorism content than cryptographic hash matching.

12.18 We do not propose recommending the use of a specific perceptual hash function. The most appropriate choice of hash function may vary depending on the media type targeted.[365]

12.19 Service providers will have the flexibility to set the threshold to determine when there is a sufficient similarity between the hash and piece of content for it to have reason to suspect the content may be terrorism content. We consider this flexibility important because it allows service providers to implement the approach that best suits their systems, and the risks on their service, while still achieving the intended outcomes. Additionally, it allows service providers to fine-tune their hash matching technology, where needed, to enable more accurate detection and to reduce the spread of terrorism content more effectively.

12.20 We propose that service providers ensure the hash matching technology is configured to strike an appropriate balance between precision and recall. In doing so, providers should consider:

- the risk of harm from terrorism content on the service, reflecting the service's risk assessment and any other information reasonably available to the provider;
- the proportion of content detected as a match by the technology that is a 'false positive';[366] and
- the effectiveness of the systems and processes used to identify false positives.

12.21 The level of false positives (including any cases arising from content being incorrectly included in a hash database) determines the potential impact of hash matching on users'

[363] Schedule 4, paragraph 13 to the Act.

[364] Global Network on Extremism and Technology: Extremists are seeping back into the mainstream: Algorithmic detection and evasion tactics on social media platforms March 2022 [accessed 16 April 2025]

[365] For example, perceptual hash functions such as pHash or Microsoft's PhotoDNA are more appropriate for images, while a perceptual hash function such as Meta's TMK+PDQF is better suited for video content.

[366] Detected content that is not target content, i.e. that is not terrorism content.

freedom of expression and privacy. The impact can be mitigated by review of detected content before action is taken in response to a match, including review by human moderators, and the operation of a complaints procedure which enables users to appeal if they believe their content has been wrongly identified as terrorism content.

12.22 Providers should ensure that the performance of the technology, and whether the balance between precision and recall continues to be appropriate, is reviewed at least every six months. This promotes transparency and appropriate record-keeping of a service provider's use of proactive technology. It also ensures service providers review and update the technology, if appropriate, based on its effectiveness and impact.

The set of hashes

12.23 The technology should use a suitable perceptual hash function to compare relevant content to an appropriate set of hashes. The service provider can choose to set up its hash list with hashes of terrorism content[367] identified by its content moderation function, hashes sourced from a person (or persons) with expertise in the identification of terrorism content, or a combination of both.

12.24 The requirements of the person with expertise are that the person has arrangements in place:

- to identify suspected terrorism content;
- to secure (as far as possible) that terrorism content is correctly identified before hashes of that material are added to its database;
- to prevent discrimination on the basis of protected characteristics in relation to identifying or assessing suspected terrorism content;
- to regularly update the database with hashes of terrorism content;
- to review cases where material is suspected to have been incorrectly identified as terrorism content and remove such hashes from the database where appropriate; and
- to secure the database from unauthorised access, interference or exploitation.

12.25 These arrangements are intended to ensure that the database provided is accurate and effectively maintained. Providers should ensure that the latest versions of any databases are regularly obtained.

12.26 At this stage, it is our understanding that there is currently not a "person with expertise" who meets the requirements set out in paragraph 12.24 in particular because the databases currently made available by third parties are not limited to hashes of terrorism content as defined in the Act but also contain hashes of other terrorist and violent extremist material. We have included the option of sourcing hashes from a "person with expertise" to allow for sector developments and the potential efficiencies of such an option, should it become available in the future. We would welcome responses on this as part of the consultation process.

12.27 The set of hashes used by a service provider may also include other hashes of terrorism content, for example hashes sourced from a person who does not meet the requirements in paragraph 12.24.

[367] As defined in section 59(8) of the Act.

12.28 Where the set of hashes includes hashes not sourced from a person with expertise meeting the requirements described in paragraph 12.24 providers should ensure that arrangements are in place:

- to secure that terrorism content is correctly identified before hashes of the material are added;
- to prevent discrimination on the basis of protected characteristics in relation to identifying or assessing suspected terrorism content;
- to regularly update the set of hashes with hashes of terrorism content identified by its own content moderation function and (where appropriate) other hashes of terrorism content; and
- to review cases where material is suspected to have been incorrectly identified as terrorism content and remove such hashes where appropriate.

12.29 Taken together, the set of hashes should reflect the range of content that amounts to an offence specified in Schedule 5 to the Act that is appropriate having regard to the risk assessment of the service and any information reasonably available to the provider about relevant content that is terrorism content on the service. This is to ensure the hash database reflects the types of visual terrorism content that might manifest on a particular service.

12.30 Providers should ensure an appropriate policy is put in place, and measures are taken in accordance with that policy, to secure any hashes from unauthorised access, interference or exploitation.

Review of detected content

12.31 Where content is detected by hash matching, providers should treat this as reason to suspect that the content may be terrorism content. [368] Detected content should be reviewed and assessed by the provider's content moderation systems and processes to determine whether it is illegal content and/or illegal content proxy (Measure ICU C1 – Having a content moderation function to review and assess suspected illegal content). If so, it should be swiftly taken down in accordance with Measure ICU C2 (Having a content moderation function that allows for the swift take down of illegal content), unless currently not technically feasible to achieve this outcome

12.32 Providers' content moderation functions are likely to include a combination of automated moderation systems and processes, and human moderation. We propose that providers should ensure that their systems and processes for the purpose of Measure ICU C1.2 are operated to ensure human moderators review and assess an appropriate proportion of detected content, having regard to:

- the degree of accuracy achieved by their automated systems and processes in use for the purposes of ICU C1.2 (taking account of data from the service's complaints procedures);

[368] In this consultation, we are proposing to introduce a measure into our Illegal content Codes of Practice for user-to-user services which recommends that providers exclude content indicated potentially to be priority illegal content from recommender feeds (see chapter 14). Under this measure, service providers should use 'relevant available information' to make this assessment. We define 'relevant available information' to include 'indicators generated by technology used on the service'. Where service providers detect content using hash matching technology in line with our proposals in this chapter, this is an 'indicator generated by technology used on the service' and should therefore be treated as 'relevant available information' for the purpose of our proposed recommender systems measure.

- the principle that content with a higher likelihood of being a false positive should be prioritised for review; and
- the importance of understanding the purpose, meaning, and context of detected content when determining whether it is terrorism content.

12.33 Human review of an appropriate proportion of detected content is an important safeguard for ensuring content is not erroneously removed as terrorism content and thereby for protecting freedom of expression.

Databases and hash matching tools

12.34 The effectiveness and accuracy of the hash matching technology is intrinsically linked to the scale and accuracy of the underlying set of hashes.

12.35 Services already using hash matching take a variety of different approaches, for example:

a) By maintaining an internal database of hashes of content previously detected on their service;
b) By accessing third-party databases, which in some cases may also perform the hash matching process on the service's behalf;
c) By layering various approaches to hash database sourcing and implementation to provide a more layered and robust response to detection of terrorism content.

12.36 The approach services currently take depends on a number of factors, including:

a) Their ability to third-party databases;
b) Their technical capacity to implement the technology; and
c) The resource they have available to dedicate to all aspects of the hash matching process (including content review, takedown and action against users).

12.37 For example, larger services with greater capacity and experience in running these systems may operate their systems entirely in-house, whereas a smaller service may outsource parts of the process to an external provider through a plug in or API.

12.38 We currently consider that service providers will need to build or develop their own internal database in order to store the hashes of terrorism content identified and removed on their service and other hashes, sourced from a third party, which they have verified to be terrorism content as set out in paragraph 12.28.

Third-party databases

12.39 Services may opt to access hashes from a third-party to supplement their own hash list.

12.40 The largest and most well-known third-party database is the industry hash sharing database (HSDB) offered by the Global Internet Forum to Counter Terrorism (GIFCT). The HSDB contains perceptual and cryptographic hashes and operates with a series of labels that help members understand the context of the content (such as content type, the terrorist entity that produced it, and its behavioural elements).

12.41 The current HSDB criteria for inclusion in the database include:

- The content is associated with a named entity on the United Nations Security Council's Consolidated Sanctions List;
- The content meets GIFCT's behavioural inclusion criteria;[369] or

[369] 1. Content is produced by a non-governmental entity. 2. Content has extremist identifiers. 3. Content has a clear core hate-based ideology. 4. Content directly advocates for violence.

- The content leads to the activation of GIFCT's Content Incident Protocol or Content Incident levels of GIFCT's Incident Response Framework.[370]

12.42 These broad criteria indicate that the hashed content in the HSDB may not directly align with content that constitutes an offence specified in Schedule 5 to the Act.

12.43 A service that chooses to use the HSDB would need to integrate it with its hash matching system, which would allow it to supplement its own database with hashes from the HSDB that it has verified to be hashes of terrorism content as defined in the Act. Services that wish to join the GIFCT and access the HSDB need to meet membership criteria.

12.44 Tech Against Terrorism hosts and operates its terrorist content analytics platform (TCAP). The TCAP is a database where terrorism content is identified, verified against the TCAP's inclusions policy[371]and classified on the TCAP. Once this process is completed, an alert is sent to the service.[372] This alert is sent to small services on a free and advisory basis.[373] Services need to review the content raised to them and proceed with content moderation. Depending on the content moderation decision, services can then choose whether to hash the content they have been alerted to on their service and add it to their own internal database.[374]

12.45 These are two third party databases that we are aware of and we would welcome further information on third party databases in response to this consultation.

Hash matching tools

12.46 There are number of open-source hash matching tools that services could choose to use. Some of these tools are free, although there are additional software, hardware and data costs (see paragraph 12.60 for more details).[375] [376]

Effectiveness at addressing risks and benefits

12.47 Hash matching technology has been deployed for some time across industry to detect terrorism content but it is not currently used by all services. If enacted, this proposed measure would significantly expand the number of service providers that use hash

[370] Global Internet Forum to Counter Terrorism Hash Sharing Database Review: Challenges and Opportunities December 2024 [accessed 25 April 2025]

[371] The inclusion policy tiers include: 1. Threat to life. 2. Crisis. 3. Designation. 4. Promotional. 5 Recommended. Terrorist Content Analytics Platform Inclusion Policy [accessed 16 April 2025]

[372] User-to-user services need to have signed up to the TCAP service with Tech Against Terrorism.

[373] Tech Against Terrorism state this is free for smaller platforms and academics based on their acceptable use policy. Tech Against Terrorism Terrorist Content Analytics Platform FAQ [accessed 20 March 2025]

[374] Tech Against Terrorism Terrorist Content Analytics Platform [accessed 20 March 2025]

[375] For example, Meta's Hasher-Matcher-Action (HMA) is a free open-source tool that allows a service to hash match images and videos. The tool then allows the service to review matches and decide what action to take. HMA can also be plugged into an internal or external database. Meta launches new content moderation tool as it takes chair of counter-terrorism NGO [accessed 16 April 2025]

[376] In addition to the HMA tool, the Altitude tool is a free open-source tool developed by Google Jigsaw and Tech Against Terrorism. The tool allows services to review matches for terrorism content identified on their service and can receive alerts via either Tech Against Terrorism's TCAP or GIFCT's HSDB. The tool then allows the service to review matches and decide what action to take. Tech Against Terrorism Altitude Content Moderation Tool [accessed 16th April 2025].

matching technology to detect terrorism content. Our analysis suggests this would deliver significant benefits.

12.48 Given the volume of content posted on user-to-user services, it is not possible for providers of the largest services to manually review a large proportion of the content they carry. There are significant limitations to the role user reporting can play in detecting terrorism content. Therefore, automated content moderation technologies such as hash matching can play an important role in the moderation of terrorism content and can materially increase the amount of such content which is detected and removed.

12.49 This is borne out by evidence from service providers that are already using hash matching. For example, between April 2022 and June 2022, 95% of the videos (of a total of 72,990 videos during this period) removed by YouTube under its violent extremism/terrorism policy were first automatically flagged.[377] Between October and December 2024, 10.5 million pieces of dangerous organisations' content was removed from Facebook. 99.4 percent of this content was identified and removed through proactive means such as hash matching.[378]

12.50 Given the harm terrorism content causes, we provisionally consider that, by improving service providers' ability to detect and remove such content, this measure would deliver significant benefits.

Impacts and Costs

12.51 This section considers the costs associated with implementing perceptual hash matching for the detection of terrorism content. Based on the discussion in paragraphs 12.11 to 12.50, we assume the measure would be deployed using an appropriate hash database and appropriate technical parameters. In addition, the proposed measure recommends the appropriate use of human review or moderation to help ensure that the hash matching technology operates accurately. We include these factors in our discussion of costs in the following paragraphs.

12.52 Many of the services that would be in scope of this proposed measure would also be within scope of the existing measure recommending the use of hash matching to detect and remove CSAM (Measure ICU C9 in the Illegal Content user-to-user Codes).[379] We expect these service providers to incur significantly lower additional costs if they are already implementing hash matching for CSAM.

12.53 The main costs are likely to consist of:

- the one-off costs associated with setting up a hash matching system to detect terrorism content, including the setting up of an internal database;[380]
- the cost of maintaining the hash matching system;
- the cost of software, hardware, and data; and

[377] YouTube Community Guidelines enforcement – Violent Extremism [accessed 16th April 2025]

[378] Facebook Community Standards Enforcement Report: Dangerous Organisations and Organised Hate [accessed 16 April 2025]

[379] We expect there to be a strong overlap between services that are at risk of CSAM, IIA and services that are at risk of terrorism content.

[380] Regardless of whether services choose an internal or third-party solution, we anticipate they will still be required to set up an internal database.

- the cost of reviewing matches, moderating content, and dealing with appeals.

12.54 There may also be other potential costs, particularly if a service decides to use hashes sourced from a third-party database to supplement its own internal systems.

12.55 The overall cost to service providers from implementing our proposed measure is likely to depend on a number of factors such as their existing systems and processes, number of users, technical complexity, and risk for terrorism content. Table 12.1 outlines illustrative total costs for a large service implementing hash matching for terrorism content with no existing hash matching systems.[381]

Table 12.1: Estimated costs of the terrorism hash matching measure for a hypothetical large service

	Estimated cost
Build cost (one-off)	£19,000 to £350,000
Maintenance costs (annual)	£5,000 to £88,000
Software, hardware and data (annual)	£100,000 to £150,000
Human moderation (annual)	£11,000 to £22,000

One-off costs to build the hash matching system

12.56 The one-off costs associated with setting up a hash matching system to detect terrorism content and store hashes will primarily consist of labour costs. It will require input from software engineers, supported by other professionals such as analysts, product or policy managers, and legal teams. The technological solution used to integrate hash matching into a service affects these development costs. The technical complexity of the service also affects these costs, as integration will be more challenging for more complex services with a multitude of functions. Larger services may have more complex operational structure, and a greater number of individuals involved in making changes to a service, which can increase the resource required to implement a new technology.

12.57 We estimate that building a hash matching system for terrorism content would take between two to eighteen months of full-time work by a software engineer. This corresponds with a one-off build cost of approximately £19,000 to £350,000.[382] We expect that smaller and less complex services are likely to incur costs towards the lower end of this range. If a service already deploys hash matching for CSAM, there may be synergies with the technology used for a hash matching system for terrorism content, which could lead to significant cost savings. This is due to the sharing of supporting architecture across different types of content and, in some cases, costs will be lower than the range we have provided.

[381] A service with seven million monthly UK users, and a global user base of seventy million users for the purposes of estimating moderation costs.

[382] This is based on our cost assumptions outlined in Annex 15.

Ongoing maintenance costs

12.58 The ongoing maintenance costs will primarily consist of labour costs associated with activities such as applying updates, reviewing the hash matching technology's performance and configuring the technology (at least every six months), ingesting new hashes from third-party databases (if used), and integrating with new service functionalities. This will require input from software engineers as well as other professional occupation staff. In line with our standard assumption, we assume that the annual maintenance costs are 25% of the one-off costs of setting up a hash matching system for terrorism content. Based on this, we estimate that the annual cost of maintaining a hash matching system for terrorism content to range from £5,000 to £88,000 depending on the technical complexity of the service.[383]

Ongoing software, hardware and data costs

12.59 The ongoing software, hardware and data costs to service providers will mainly consist of the costs associated with operating their own hash databases, accessing hash databases and utilising API solutions from third-party organisations where relevant. These costs generally tend to scale with a service's user base and the volume of content that must be checked and hashed. Therefore, we expect service providers with a large user base will incur greater costs than a smaller service. Software and hardware costs would be incurred in addition to the initial build cost and general ongoing maintenance costs. A service provider would need to set up its own infrastructure and database prior to integrating external software.

12.60 An example of a third-party hash matching tool is Meta's Hasher-Matcher-Actioner (HMA) which is a free open-source trust and safety tool that can be used in conjunction with a hash database and can hash images and videos.[384] While the tool is free, a service provider would need additional software or hardware to run HMA (such as a cloud API or acquiring their own computing infrastructure). We estimate a large service would incur annual costs of £100,000 to £150,000 to hash and compare eighty-seven million pieces of content per day.[385] Smaller services would require far fewer queries and therefore would be able run HMA at a significantly lower cost than the range we have estimated. The largest services would incur higher costs due to their complexity and greater volumes of content. We consider that the largest services are more likely to develop their own internal hash matching tools as they are more likely to have the necessary financial and technical resources and could develop to their individual requirements.

Ongoing content moderation costs

12.61 The ongoing content moderation and appeals costs will depend on the technical set-up of the service, the quantity of content uploaded to the service, and the quantity of terrorism content present on the service. These costs will mainly consist of the labour costs associated with reviewing content that is flagged by the hash matching technology, moderating the content and dealing with appeals. We therefore assume that the annual moderation cost will depend on the number of matches that are detected on the service by

[383] This uses our standard assumption that annual maintenance costs are equal to 25% of build costs as set out in Annex 15.

[384] Meta launches new content moderation tool as it takes chair of counter-terrorism NGO [accessed 16 April 2025]

[385] We define a large service as a service with more than seven million monthly active UK users.

the hash matching technology. We generally expect the costs of human moderation to scale with the amount of content on a service and to be higher for services that have more terrorism content. Therefore, we expect that the costs of human moderation, including those incurred by smaller service providers, will correlate with the extent of harm that exists on a service, and are therefore likely to be proportional to the benefits of our measures.

12.62 We estimate for a hypothetical large service with seventy million monthly users globally, of which seven million are in the UK, moderation costs could be between £11,000 to £22,000 per year. This is the cost of human moderation needed to review and action around 2,600 pieces of content flagged by the hash matching system.[386]

12.63 Once service providers have determined content is terrorism content, they will need to take appropriate action to take down this content. Service providers will also need to operate an appeals process, so that users can flag if their content has been erroneously removed. We expect the upfront costs associated with these steps to already be covered by existing code measures (content moderation and reporting and complaints) which apply to all services; however, we acknowledge that our proposed measure may result in more content being identified for takedown, and potentially more appeals.

Third-party solutions

12.64 Services may opt to access a third-party solution for the hash matching technology or database. For example, services could use the HSDB offered by GIFCT as a source of additional hashes, as discussed in paragraph 12.43. A service that chooses to use the HSDB would need to verify and integrate it with its hash matching system which would allow it to supplement its own database, and review hash matches on its service. Services that wish to join the GIFCT and access the HSDB need to meet membership criteria. There are also costs associated with membership of this organisation, which scale with the size of the service.

Wider market impacts

12.65 Hash matching is a relatively mature technology that is already being used by many of the largest services (and some smaller services). There may be benefits for the market should there be a wider deployment of hash matching technology. It is possible that this proposed measure could increase demand for hash matching technology which drives innovation and competition leading to service providers having a wider range of options when they consider how to implement the measure. In the longer term this could result in lower costs and an improvement in the effectiveness of hash matching systems.

Rights assessment

12.66 This section considers the proposed measure's impact on users' rights to freedom of expression and to respect for private and family life. As explained in chapter 3, restrictions on those rights must be necessary and proportionate. Our assessment of the measure's adverse impacts on users' rights is therefore to be balanced against its contribution to removing terrorism content. The UK Parliament has legislated for terrorism content to be

[386] These estimates are based on data provided in Snap's 2024 Digital Services Act report and global transparency reports from H1 and H2 2023 [accessed 22 May 2025]. We calculated from these data that on average there are 0.00004 items of terrorism content uploaded per user of a service and 0.0013 moderators needed per piece of content actioned. Additional cost assumptions are outlined in Annex 15.

designated as 'priority illegal content' under the Act, requiring service providers to use systems and processes designed to minimise the length of time for which it is present. This reflects the very substantial public interest that exists in measures that reduce its prevalence and dissemination online.

Freedom of expression

12.67 Detecting and removing terrorism content is in the legitimate interests of national security, public safety and acts to prevent crime by deterring users from posting such content and preventing other users from encountering it. Combatting terrorism is unarguably a pressing social need. Paragraphs 12.1 to 12.4 set out the harm to users and wider society caused by terrorism content and the measure's effectiveness in detecting such content so that it can be removed.

12.68 Interference with users' freedom of expression arises where content is wrongly taken down by the systems and processes implemented in accordance with this measure. Adverse impacts could arise in relation to the most highly protected forms of content, such as religious and political expression, and in relation to kinds of content the Act seeks to protect, such as content of democratic importance or journalistic content. Such content may be particularly vulnerable to being incorrectly detected as a false positive when hash matching for terrorism content. This is because the purpose, meaning and context of content is important in making judgements regarding potential terrorism content. As we explain in the Illegal Content Judgements Guidance Chapter 2 'Terrorism', it is not an offence to portray terrorism (for example, in a clip from a film), or to report on terrorism (for example, as news or current affairs). The purpose of the person posting content, including whether they are a member of a proscribed organisation, is relevant to determining whether the content amounts to a terrorism-related offence.[387] There are public interest reasons for organisations and individuals, such as law enforcement authorities, anti-terrorism organisations, academic researchers, journalists and human rights organisations to post content about proscribed organisations.[388]

12.69 The risk of false positives has a particular impact on users more likely to post content that is similar to known terrorism content. Similarities could exist across the type of content or the medium by which it is shared. We understand activists may be apprehensive about documenting legitimate issues and these getting surfaced in perceptual hashing matches.[389]

12.70 There is also a risk of hash databases disproportionately focusing on Islamist terrorist groups and material, notwithstanding the wider range of terrorist ideologies, especially extreme right-wing terrorism. This is because the majority of international terrorist organisations proscribed by the UK Government, by other states, and by the United Nations are Islamist. This reflects recent geopolitical circumstances. However, the number of extreme right-wing terrorist organisations being added to the list of proscribed terrorist organisations under UK law is increasing. As a consequence, incorrect identification of content as terrorism content may disproportionality affect Muslims if content of a religious or political nature is conflated with that related to terrorism.

[387] Ofcom Illegal Content Judgements Guidance, paragraphs 2.4 and 2.5.

[388] Ofcom Illegal Content Judgements Guidance, paragraph 2.48.

[389] Electronic Frontier Foundation, 2019. Caught in the Net: The impact of "extremist" speech regulations on human rights content [accessed 16 April 2025]

12.71 Risks of the incorrect detection of content as terrorism content can be mitigated by the safeguards which we are proposing in relation to the set of hashes used and the technical configuration of the hash matching technology. For example:

- The measure proposes that arrangements should be in place to secure that terrorism content is correctly identified before hashes of that material are added to the set of hashes used and to review any cases where material is suspected to have been incorrectly identified as terrorism content. The provider should ensure that the arrangements in place do not plainly discriminate on the basis of protected characteristics in identifying or assessing content.
- The hash matching technology should be configured to strike an appropriate balance between precision and recall. In striking this balance providers should have regard to, among other things, the proportion of content detected as a match by the technology that is a 'false positive', and the effectiveness of the systems and processes used to identify false positives. This balance should also be reviewed every six months.

12.72 The measure also provides for the further moderation of content that is detected by hash matching, in accordance with Measures ICU C1 and C2. This should include human moderation of an appropriate proportion of detected content, having regard to the accuracy achieved by the provider's automated systems/processes, the principle that content with a higher likelihood of being a false positive should be prioritised for review, and the importance of understanding the purpose, meaning and context of detected content when determining whether it is terrorism content. This is an important mitigation given the context-specific nature of terrorism content.

12.73 Other measures also act as safeguards for users' freedom of expression; in particular, providers are required to take appropriate action in response to certain complaints, including appeals by UK users who have generated, uploaded or shared content on a service which has been taken down on the basis that it is illegal content.

12.74 Interference with users' freedom of expression may also arise where providers take action against those users (such as banning an account) because the user has been detected as sharing terrorism content. We consider that impact in chapter 17. The safeguards included in this measure to protect users' freedom of expression would limit the risk that action is taken against users on the basis of false positives.

12.75 Overall, we acknowledge that the measure involves interference with users' rights to freedom of expression where content is incorrectly detected as terrorism content, but we consider the interference to be limited, with safeguards in place, and proportionate to the measure's aim of reducing the prevalence and dissemination of terrorism content.

Privacy and data protection

12.76 Hash matching involves the automated processing at scale of content uploaded, generated or shared by users. This will often include the processing of personal data. Services will need to ensure that the automated processing involved in hash matching and all associated processing (such as human review of detected content) are carried out in compliance with data protection law. The Information Commissioner's Office (ICO) has provided guidance on Content moderation and data protection.[390]

12.77 Where the automated processing involved in hash matching is carried out in compliance with data protection law, it should have minimal impact on users' privacy. Review of detected content by human reviewers may have a more significant impact on users' privacy. However, since the measure will only apply to content communicated publicly, the user sharing the content should have a reduced expectation of privacy in connection with the content.

12.78 The same measures that act as safeguards for users' right to freedom of expression also safeguard users' privacy (including the protection of personal data) in that they promote compliance with the data protection principles of accuracy, fairness and transparency, and assist users to exercise their rights under data protection legislation.

12.79 Interference with users' privacy could also result from action taken against users by providers because the user has been detected as sharing terrorism content. We consider that impact in chapter 17. The safeguards included within this measure would limit the risk that action is taken against users on the basis of false positives.

12.80 Overall, we consider that the impact on users' rights under Article 8 of the European Convention on Human Rights (ECHR) is proportionate to the measure's aim of reducing the prevalence of and dissemination of terrorism content.

Which providers should implement this measure

12.81 Our analysis in this chapter shows that the proposed measure would deliver significant benefits, but that it would also impose significant costs.

12.82 We are proposing to apply the measure to providers of user-to-user services that enable regulated user-generated content in the form of photographs, videos or visual images (whether or not combined with written material) to be generated, uploaded or shared and:

a) are large[391] services that are at medium or high risk of terrorism content; or
b) are at high risk of terrorism content, and (1) have more than 700,000 monthly active UK users; or (2) are file-storage and file-sharing services.

12.83 The reach and the size of large services means that the impact of terrorism content circulating on these services could be very significant.

12.84 Many of the high risk services with between 700,000 and seven million users are likely to be in scope of the CSAM hash matching measure and therefore would likely incur a reduced incremental cost. As set out in paragraphs 12.1-12.4 and the Register of Risks, there is significant potential of harm for users from encountering terrorism content. Restricting this measure to large services risks the proliferation of terrorism content on smaller services. Bearing this in mind and considering the severity of the harm caused by terrorism content, we therefore provisionally conclude that it is proportionate to apply this measure to services with between 700,000 and seven million users where they are at high risk of terrorism content.

[390] The ICO has also provided guidance on Automated decision-making and profiling.

[391] A large service is one with more than seven million monthly UK users.

12.85 File-storage and file-sharing services are critical to enabling terrorist groups to disseminate their online content. Research conducted in 2021 by the Institute for Strategic Dialogue identified a large cache of terrorist content (in the region of 2.2 terabytes) produced by the Islamic State of Iraq and Syria stored on a file hosting application.[392] Research conducted by Tech Against Terrorism found that, between 25 November 2020 and 19 January 2023, file sharing platforms represented over half of the platforms on which they identified terrorist content (106 of 187 platforms).[393]

12.86 We consider that certain file-storage and file-sharing services are likely to be at especially high risk of terrorism content. These services can host large volumes of terrorism content which can be disseminated on other services, even when the file-storage and file-sharing service is small.[394] Smaller file-sharing and file-storage services would have fewer financial resources to implement hash matching so this measure would likely have a greater relative burden. Most of these services are likely to be in scope of the CSAM hash matching measure and therefore would likely incur a reduced incremental cost. We provisionally conclude that this measure applies to all providers of file-storage and file-sharing services that are at high risk of terrorism content.

12.87 We do not consider it proportionate to recommend the measure for high risk services with 700,000 monthly UK users or fewer, that are not file-storage and file-sharing services. These services are unlikely to be in scope of other hash matching measures. The relative cost of implementing hash matching will therefore be substantial for these services while the smaller user numbers mean there is a reduced risk of harm from users encountering terrorism content.

Provisional conclusion

12.88 Taking account of the severity of the risk of harm to users and other individuals associated with the dissemination of terrorism content online, our provisional view is that it is proportionate to recommend the use of perceptual hash matching for terrorism content in our Illegal Content Codes in relation to providers of user-to-user services that enable regulated user-generated content in the form of photographs, videos or visual images (whether or not combined with written material) to be generated, uploaded or shared and:

- are large[395] services that are at medium or high risk of terrorism content; or
- are at high risk of terrorism content, and:
  - (1) have more than 700,000 monthly active UK users; or
  - (2) are file-storage and file-sharing services.

12.89 This measure would make it more difficult to disseminate or share terrorism content and will have a direct impact on the availability and reach of terrorism content to adults and children across in scope user-to-user services. This will in turn result in a decrease in exposure of users of all ages to content which has the potential to encourage or promote terrorism. We consider this measure to be appropriate on the basis that the cost is proportionate considering the significant harm terrorism content causes.

[392] Institute for Strategic Dialogue, The Cloud Caliphate: Archiving the Islamic State in Real-Time May 2021 [accessed 16 April 2025]
[393] Tech Against Terrorism Patterns of Online Terrorism Exploitation April 2023 [accessed 16 April 2025]
[394] Tech Against Terrorism response to November 2023 Consultation, p.2. [accessed 16 April 2025]
[395] A large service is one with more than seven million monthly active UK users.

13. Proposal to extend perceptual hash matching for child sexual abuse content

Summary

The Codes already require some services to use hash matching to identify known CSAM material. In this consultation, we are proposing to extend this requirement to high-risk services whose primary purpose is hosting or disseminating regulated pornographic content.

Our proposals

Number	Existing measure	Proposal: Who should implement this
Amendment to ICU C9	Providers should ensure that hash-matching technology is used to detect and remove child sexual abuse material (CSAM).	Extending the measure to providers of user-to-user services which are **high risk** of image-based CSAM and where the principal purpose of the service is the hosting or dissemination of regulated pornographic content.

Consultation questions

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

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

Detail of our proposal

13.1 In the December 2024 Illegal Harms Statement we included a provision in our codes of practice setting out that providers of the following types of service should use perceptual hash matching to detect image-based CSAM:

- Large services that are at medium or high risk of image-based CSAM.
- Services that are at high risk of image-based CSAM and (a) have more than 700,000 monthly active United Kingdom users or (b) are file-storage and file-sharing services.[396]

13.2 We are now proposing to widen the scope of the perceptual hash matching measure such that all services whose principal purpose is the hosting or dissemination of regulated pornographic content who are at high risk of image-based CSAM, regardless of size, should implement the measure.

13.3 This aligns the measure with our proposal for hash matching for intimate image abuse (IIA), where we recommend that providers of user-to-user services which are high risk of intimate image abuse and whose principal purpose is the hosting or dissemination of

[396] Ofcom, 2024. Illegal content Codes of Practice for user-to-user services, ICU C9 [accessed 13 June 2025].

regulated pornographic content should implement hash matching for intimate image abuse content.

Benefits and effectiveness at addressing risks

13.4 We consider it appropriate to extend our CSAM hash matching measure in line with the measure for IIA. In the same way as for intimate image abuse content, pornography services are known to be a high-risk space for CSAM. For example, Protect Children Finland found that 31.7% of offenders surveyed had encountered CSAM on a pornography service, making it the most common type of online service identified in the survey for where CSAM was encountered on the clear web.[397] This also indicates that individuals may inadvertently stumble across this material, which can be a source of considerable distress for users and can also act as a gateway to further offending.

13.5 Pornography services are known within industry to be a high-risk area for CSAM, with the Internet Watch Foundation launching a chatbot alongside the Lucy Faithfull Foundation and Aylo to help potential offenders address and stop their behaviour on the Pornhub service.[398] The chatbot was displayed 2.8 million times between March 2022 and August 2023 in response to potential searches for child sexual abuse material, indicating that users are seeking this content on pornography services in high volume.[399]

13.6 We assess that a number of pornography services hosting CSAM fall under the current 700,000 UK user threshold. Confidential evidence seen by Ofcom showed that there were up to [✂] pornography services found hosting CSAM in 2024, with the actual number of such services hosting CSAM likely to be higher.[400] [✂] A sample taken by Ofcom of these services found that they had fewer than 700,000 monthly UK users per month in recent months, and we assess that this is likely to be the case for many of the services hosting CSAM that were identified. Applying the measure to all services whose principal purpose is the hosting or dissemination of regulated pornographic content would capture these services.

13.7 Tackling CSAM on pornography services tackles the issue of content depicting abuse of post-pubescent children (16-17 years old), which are found in high volume on such services. For example, in 2021 the hosting provider Serverel received 66,824 removal notices from the Canadian Centre for Child Protection (C3P)'s Project Arachnid for hosting post-pubescent CSAM. C3P state that their data suggests that 'many websites' using Serverel hosting services are 'ephemeral adult content websites that host post-pubescent material among legal adult content'. C3P also suggests that this type of abuse is likely to circulate on pornography services due to perpetrators' 'plausible deniability' of the age of the child in the content, and individuals' own 'lack of understanding about what constitutes CSAM' and the consequences of disseminating this content.[401]

[397] Protect Children, 2024. Tech Platforms Used by Online Child Sexual Abuse Offenders [accessed 21 May 2025].

[398] Lucy Faithfull Foundation, 2022. Stop It Now, Internet Watch Foundation and Pornhub launch first of its kind chatbot to prevent child sexual abuse [accessed 21 May 2025].

[399] The University of Tasmania (Scanlan, J. et al.), 2024. reThink Chatbot Evaluation [accessed 21 May 2025].

[400] Email correspondence with [✂], 27 March 2025.

[401] Canadian Centre for Child Protection, 2021. Project Arachnid: Online availability of child sexual abuse material. [accessed 12 May 2025].

Impacts and costs

13.8 As described in our December 2024 Illegal Harms Statement[402], we expect the main costs of the measure are likely to consist of:

- one-off costs related to building a hash-matching system;
- the cost of maintaining a hash-matching system;
- the cost of software, hardware, and data; and
- the cost of reviewing matches, moderating content, and reporting CSAM to external bodies.

13.9 We expect the main risks of the measure are:

- Incorrect detection of content as CSAM;
- security compromises; and
- potential biases in hash databases.

13.10 This is explored fully in the December statement at paragraphs 4.103 to 4.140.

Rights assessment

13.11 The rights assessment considers the measure's impacts on user's rights under Articles 8 and 10 of the European Convention on Human Rights ('ECHR').[403]

13.12 Overall, we consider that the measure's impacts on users' rights under Articles 8 and 10 ECHR are proportionate to the measure's aim of reducing the prevalence and dissemination of CSAM. To the extent that our proposals mean that additional users (and their rights) will be impacted, we consider that impact to be proportionate to the benefits provided by the measure in reducing the prevalence of CSAM on pornography services with less than 700,000 UK users.

13.13 For the full rights assessment for this measure please see the December 2024 Statement, paragraphs 4.141-4.177.[404]

13.14 Some details of the assessment with regards to freedom of expression and privacy are set out below.

Freedom of expression and association

13.15 As set out in the December 2024 Statement, the design of the measure includes a number of safeguards to protect users' freedom of expression, directed at the set of hashes used and the technology's technical configuration. The measure also sets out processes designed to further reduce the impacts of false positives detected by the technology.

[402] Ofcom, 2024. Volume 2 Service Design and User Choice [accessed 13 June 2025].

[403] Article 10 of the ECHR sets out the right to freedom of expression; Article 8 of the ECHR sets out the right to respect for individuals' private and family life. Restrictions on those rights must be necessary and proportionate.

[404] Ofcom, 2024. Volume 2 Service Design and User Choice [accessed 13 June 2025].

13.16 Full discussion of freedom of expression impacts and safeguards are set out in the December 2024 Statement, paragraphs 4.148-4.162.[405]

Privacy and data protection

13.17 Removing CSAM is vital for protecting the right to privacy of victims of child sexual abuse and also helps to protect their personal data.

13.18 As set out in the December 2024 Statement, hash matching involves the automated processing at scale of images and videos uploaded, generated, or shared by users. This will often involve processing of personal data.

13.19 We are satisfied that the processing required by the measure can be carried out in accordance with data protection law.

13.20 For further detail of data protection impacts and safeguards see the December 2024 Statement, paragraphs 4.163-4.168.[406]

13.21 In terms of users' right to privacy, we consider that, where service providers ensure that the automated processing involved in hash matching is carried out in compliance with data protection law, that processing should accordingly have a minimal impact on users' privacy.

13.22 Some aspects that may have a more significant impact on users' privacy are:

- Review of some detected content by human moderators;
- The requirement to report CSEA content which is not otherwise reported to the NCA by providers of regulated U2U services[407], especially where this report may be a false positive[408],[409];
- Action taken against users by service providers based on a false positive[410].

13.23 The same measures that act as safeguards for users' right to freedom of expression are relevant, as they tend to promote compliance with the data protection principles of (in particular) accuracy, fairness and transparency, and to assist users to exercise their rights under data protection legislation.

13.24 For the full privacy assessment and safeguards please see the December 2024 Statement, paragraphs 4.169-4.176.[411]

[405] Ofcom, 2024. Volume 2 Service Design and User Choice [accessed 13 June 2025].

[406] Ofcom, 2024. Volume 2 Service Design and User Choice [accessed 13 June 2025].

[407] This requirement on service providers is set out in the Online Safety Act s.66 (not yet in force).

[408] The measure sets out that service providers should take into account the importance of minimising the reporting of false positives to the National Crime Agency or a foreign agency: Illegal Content Codes of Practice for User-to-user Services, ICU C9.12 (d) [accessed 13 June 2025].

[409] ICO response to November 2023 Consultation, p.13.

[410] At this stage, the Codes do not provide for such action to be taken, and this would be a matter for the service provider. It will be important for providers to take account of these potential impacts when designing their safety policies. The safeguards included within the measure, and in particular users' rights to complain about such action, also help to limit this risk.

[411] Ofcom, 2024. Volume 2 Service Design and User Choice [accessed 13 June 2025].

Provisional conclusion

13.25 There is a strong argument in principle for applying the measure to all the proposed services regardless of size. We set a threshold for user numbers in our December 2024 Statement to mitigate the risk of overwhelming third-party providers of hash databases. We have subsequently received reassurance from third-party providers that they are able to meet demand for providers seeking to implement the measure on their service. We have also identified further compliance options for service providers since the measure's publication.

13.26 Given the high volume of CSAM hosted on pornography services, the severity of the harm this causes, and to introduce consistency with the proposed hash matching measure for intimate image abuse content, we consider it proportionate to propose this change to the CSAM hash matching measure.

13.27 We assess that extending this measure will have considerable impact on reducing the proliferation of CSAM online, particularly for known CSAM depicting post-pubescent victims which may not otherwise be recognised by platform users or a service's moderators.

13.28 Given the availability of low- or no-cost options to implement the measure, the increase in available options for implementation since the original measure was proposed, and our existing assessment of potential rights impacts and safeguards, we assess that the proposed extension of the measure is proportionate.

14. Recommender Systems

Summary

Recommender systems provide a curated feed of recommended content for a user. Recommender systems are widely used and benefit users by serving them with content that they are likely to find relevant or interesting. However, recommender systems can also amplify illegal content – such as hate, terrorism or suicide content, and content from hostile states trying to interfere with UK elections. This can have a substantial impact, making such content reach a large audience in a short space of time.

To combat this risk, we are proposing that, where there are indicators (for example, content labels or tags) that a piece of content might be illegal, service providers should exclude it from recommender systems unless and until their content moderation teams have reviewed it. Where content moderation teams judge that the content is illegal, it should then be removed from the service. Where they judge that it is lawful and allowed on the service they can reinstate it in their recommender systems. We propose this measure should apply to services with a content recommender system at medium or high risk of hate, terrorism, suicide or foreign interference offences.

This proposal builds on a measure in our Protection of Children Codes, which says that services should exclude content indicated potentially to be primary priority content from the recommender feeds of children.

Our proposals

Number	Proposed measure	Who should implement this
ICU E2	Providers should design and operate their recommender systems to ensure that content indicated potentially to be certain kinds of priority illegal content (relevant illegal content) is excluded from the recommender feeds of users.	Providers of user-to-user (U2U) services that have a content recommender system and are medium or high risk for at least one kind of the following kinds of illegal harms: hate, terrorism, suicide, or foreign interference offences (FIO) (relevant harms).

Consultation questions

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

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

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

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

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

Definition: Relevant illegal content

This term is used throughout the chapter to refer to content associated with the priority illegal harms for which we have noted recommender systems as being a risk factor. These are found in our Illegal Harms Register of Risks and our Risk Assessment Guidance. In this case, this term refers to the harms of hate, terrorism, suicide, and foreign interference offences (FIO), or the 'relevant harms'. In this chapter, 'illegal content' should always be understood to mean relevant illegal content.

What are recommender systems?

14.1 Content recommender systems are algorithmic systems that determine the relative ranking of an identified pool of content from multiple users on content feeds. This includes regulated user-generated content such as images, videos, posts, and livestreams. Content is recommended based on factors that the system is designed to account for, such as popularity of content, characteristics of a user, or predicted engagement.

14.2 Our definition does not include recommender systems that suggest content to users in direct response to a search query, network recommender systems that suggest other users to follow or groups to join (such as network expansion prompts), or product recommender systems. Nor do we include content recommender systems that only recommend a users' own content from their own private inventory where this is not capable of being encountered by another user. See Annex 17 for more information and a detailed definition.

What risks do recommender systems pose?

14.3 As noted in the Illegal Harms Register of Risks (the Register),[412] if illegal content is posted to a user-to-user service and is not quickly detected and removed by existing moderation systems, recommender systems may spread that content further and expose more users to it before it is removed from the service. This includes both accidental or unwanted exposure to such content, as well as increasing the frequency and severity of exposure for users who actively seek it out.

Harm to users

14.4 In the Register, we note that recommender systems are a risk factor for exposing users to several types of illegal content – in this case terrorism, hate, suicide, and foreign interference offence (FIO) content – often without the user wishing to see it.

14.5 Recommender systems can expose users to greater volumes of such content, which can become more severe over time, resulting in in them experiencing a 'filter bubble' (an echo chamber of thematically homogenous content or content with a unified theme) or 'rabbit

[412] Ofcom, Illegal Harms Register of Risks, 2024. [accessed 13 June 2025]

hole' (content of increasing thematic intensity) effect.[413] [414] This can also leave them vulnerable to confirmation bias.[415] This risk is also exacerbated by users being drawn towards incendiary content under some platform designs.[416] A user's, or similar users', positive engagement signals (such as clicks and watch time) may be inferred by the recommender system as relevant to these users, irrespective of whether the user has a genuine interest in that content.[417]

14.6 In the Register, we provided evidence as to how users' implicit and explicit engagement with harmful content can lead to users being recommended more of such content.[418] [419] We did so to highlight the importance of design choices when managing risk of harm to users. For example, the signals that are used in recommender systems and the relative level of influence (i.e., weighting) assigned to them, can influence how the system learns about user behaviour and recommends content.

14.7 Some recommender system configurations may carry an increased risk of amplifying harmful and illegal content (for example, by being optimised for predicted engagement).[420] Testing and safety-oriented evaluation of recommender systems can also play a role in assessing the level of a risk to users on a service. Insufficient testing and not observing safety-oriented metrics could result in the deployment of design choices that carry an increased risk of illegal content amplification.[421]

Recommender design and adversarial use

14.8 Another major risk posed by recommender systems is adversarial usage or 'gaming'. The design of recommender systems affects how easily they can be exploited by bad actors that seek to manipulate, skew, or bias the outputs of the system. For example, a system that assigns high importance to likes and reshares could algorithmically amplify content that receives such signals. Such design can be vulnerable to coordinated 'gaming' attacks, in which large numbers of accounts post illegal content and then like and share the content at scale in a coordinated fashion.

[413] In the May 2024 Consultation on Protecting Children from Harms Online (May 2024 Consultation), we defined a 'rabbit hole' as 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; Ofcom, Protecting children from harms online - Volume 3: The causes and impacts of online harms to children, 2024 [accessed 13 June 2015]

[414] RUSI (Reed, A., Whittaker, J., Votta, F. and Looney, S.), 2019. Radical Filter Bubbles: Social Media Personalisation Algorithms and Extremist Content [accessed 23 April 2025].

[415] Whittaker, J., Looney, S., Reed, A. and Votta, F., 2021. Recommender systems and the amplification of extremist content, Internet Policy Review, 10(2). [accessed 23 April 2025].

[416] Munn, L., 2020. Angry by design: Toxic communication and technical architectures. *Humanities and Social Sciences communications*. 7 (53) [accessed 23 April 2025].

[417] Munn, L., 2020.

[418] RUSI (Reed, A., Whittaker, J., Votta, F. and Looney, S.), 2019 Samaritans and Swansea University, 2022. How social media users experience self-harm and suicide content. [accessed 13 June 2025]

[419] We recognise this evidence tends to focus on content which is harmful but legal. However, this tends to be a result of such studies not differentiating content on the basis of legality but rather harm. As such, this evidence is likely to capture both some illegal, and some legal (but harmful) content. We consider that illegal content that is yet to be detected and removed from the service is likely to be disseminated in a comparable way as harmful content, even if it is not illegal.

[420] Munn, L., 2020.

[421] Global Internet Forum to Counter Terrorism, 2021. Content-Sharing Algorithms, Processes, and Positive Interventions Working Group: Part 1. [accessed 23 April 2025].

14.9 Certain users may explicitly create and upload content that intends to elicit user engagement. Such content often has high shock value and is designed to be deliberately inflammatory and sensationalist. If content of such nature is being uploaded in relatively large volumes, and a large number of users are engaging with that content, then there is a risk of that content being amplified further and more users potentially experiencing exposure bias.[422] This could increasingly result in users showing a preference for illegal content where they have been exposed to it as a result of the processes described above.

14.10 This is especially likely within online communities that share 'edgy' humour or 'trashposting'.[423] 'Outrage-inducing' content in such communities is noted in the Register as increasing recommendation of hate content.[424] These groups share content with notably hateful and terroristic themes. These communities purposely create borderline content that pushes the limits of legality to avoid their content being removed from a service, but which causes sufficient controversy to drive engagement. As suggested in paragraphs 14.5-14.7, engagement with such borderline (but legal) content can eventually lead to similar (but ultimately illegal) content being recommended to users who the recommender system determines are potentially more likely to find it of interest.

14.11 Recommender systems can also be vulnerable to foreign interference operations. In the Register, we identified how perpetrators may take advantage of engagement-focused systems by acting in a coordinated fashion to generate high volumes of explicit feedback to artificially inflate the dissemination of specific posts. By artificially inflating engagement, recommender systems may then be more likely to promote this content to users. An example of such planned manipulation can be found in analysis of documents from the Social Design Agency (SDA), which indicated a desire to maximise user exposure to its activity through a range of tactics including comments via social media accounts and use of bots to spread narratives, an effort commonly known as 'Doppelganger'.[425] [426]

Legal framework

14.12 The proposed measures are founded on service providers' existing duties under the Online Safety Act 2023 (the Act) and expand on existing recommender systems measures as set out in the Illegal Content User-to-User Codes and the Protection of Children User-to-User Codes.

14.13 Part 3 of the Act places duties on providers of regulated services. These include duties set out in section 10 of the Act that require providers of regulated user-to-user services to take or use proportionate measures relating to the design or operation of the service to (among

[422] Thorsten Krause, Daniel Stattkus, Alina Deriyeva; Jan Heinrich Beinke; Oliver Thomas, Wirtschaftsinformatik 2022 Proceedings. Internationale Tagung Wirtschaftsinformatik (WI-2022), February 21-23, Nürnberg, Germany, AISeL, 2022. Wirtschaftsinformatik 2022 Proceedings | Wirtschaftsinformatik | Association for Information Systems [accessed 23 May 2025].

[423] This is also known as 'shitposting', which is defined as "the act of throwing out huge amounts of content, most of it ironic, low-quality trolling, for the purpose of provoking an emotional reaction in less Internet-savvy viewers. The ultimate goal is to derail productive discussion and distract readers"; Evans, R., 2019. Shitposting, Inspirational Terrorism, and the Christchurch Mosque Massacre - bellingcat [accessed 23 April 2025].

[424] Munn, L., 2020.

[425] US Department of Justice, 2024. Affidavit in Support of Seizure Warrant, pp. 25, 129-130.

[426] Conduct of this may, depending on the facts, fulfil the elements of the foreign interference offence. Full details are set out in chapter 14 of Ofcom's Illegal Content Judgements Guidance.

other things)[427] prevent individuals from encountering priority illegal content, effectively mitigate and manage the risk of the service being used for the commission or facilitation of a priority offence, and effectively mitigate and manage the risks of harm to individuals as identified in a service's most recent illegal content risk assessment.[428]

14.14 Priority illegal content is content that amounts to an offence specified in schedules 5, 6 or 7 to the Act (which includes relevant illegal content).[429]

Relationship with other measures

14.15 Ofcom is prioritising addressing the risks recommender systems pose, and we have already introduced measures which help to address these risks.

14.16 In our Illegal Content User-to-user Codes,[430] we recommend that certain service providers collect safety metrics during on-platform testing of proposed or actual recommender systems (ICU E1).

14.17 In our Protection of Children User-to-user Code,[431] we recommend that service providers at risk of certain types of content harmful to children should exclude content indicated potentially to be primary priority content (PPC)[432] and exclude or give a low degree of prominence to content indicated as potentially being priority content (PC)[433] and non-designated content (NDC)[434] in the recommender feeds of child users (PCU E1 and E2, respectively).

14.18 We also recommended that providers of large services enable child users to give negative feedback on recommended content, and for service providers to design and operate recommender systems to ensure negative feedback is taken into account for the purpose of reducing the likelihood of that user encountering similar content (PCU E3).

Explanation of the measure

14.19 Existing measures recommend that where a service provider has reason to suspect that content may be illegal, the provider should review that content and either make an illegal content judgement or a judgement of whether the content breaches its terms of service. Where this judgement concludes that there are reasonable grounds to infer that the

[427] Section 10(2) of the Act.

[428] For brevity, in this section we refer to 'users' rather than 'UK users'. However, for the avoidance of doubt the proposed measure only relates to the design or operation of a service in the UK or as it affects UK users (as defined in section 227(1) of the Act), consistent with Schedule 4 to the Act, paragraph 11.

[429] Section 59(1) of the Act.

[430] Ofcom, Illegal content Codes of Practice for user-to-user services, 2025. [accessed 13 June 2025]

[431] Ofcom, Protection of Children Code of Practice for user-to-user services, 2025. [accessed 13 June 2025]

[432] PPC refers to pornographic content, and content which encourages, promotes, or provides instructions for suicide; self-harm; or an eating disorder or behaviours associated with an eating disorder.

[433] PC refers to content that is abusive on the basis of race, religion, sex, sexual orientation, disability or gender reassignment; content that incites hatred against people on the basis of race, religion, sex, sexual orientation, disability or gender reassignment; content that encourages, promotes, or provides instructions for serious violence against a person; bullying content; content which depicts serious violence against (or graphically depicts serious injury to) a person or animal (whether real or fictional); content that encourages, promotes, or provides instructions for stunts and challenges that are highly likely to result in serious injury; and content that encourages the self-administration of harmful substances.

[434] NDC refers to content that is not PPC or PC but presents a material risk of significant harm to an appreciable number of children in the UK.

content is illegal, or that it is an illegal content proxy, the provider should remove the content from its service, unless it is currently not technically feasible to achieve this outcome.

14.20 This measure seeks to address risks associated with content being recommended to users before any such judgement has been made. It proposes that, where there are indicators that content is potentially illegal, the service provider should exclude the content from recommender feeds unless, and until, the service has taken content moderation action.[435] We do not propose that the content be removed from the service entirely unless (and until) the provider has reviewed and assessed the content and identified that the content is illegal content and/or illegal content proxy.

14.21 Service providers should determine the relevant available information[436] that can be used to indicate if a piece of content is potentially illegal. Their recommender systems should then be designed and operated in a way to take appropriate account of that information, and to exclude content indicated potentially to be illegal content from users' recommender feeds. This means protecting users from content that is indicated potentially to be illegal, rather than waiting for content to be confirmed as illegal or illegal content proxy through content moderation.

14.22 Service providers should take at least the following steps to exclude content indicated potentially to be illegal content from users' recommender feeds:

- Identify what relevant available information exists. We discuss the definition of 'relevant available information' in paragraph 14.25.
- Design the recommender system so as to take appropriate account of that relevant available information.
- Ensure the content recommender system operates so that content indicated potentially to be illegal content is excluded from users' recommender feeds.

14.23 We expect service providers to configure their recommender system to use a combination of indicative signals available to them to determine whether content is potentially illegal.

[435] This measure applies to illegal suicide and hate content. Content which is legal but which encourages, promotes or promotes instructions for suicide is primary priority content that is harmful to children (section 61(3) of the Act) (PPC). And content which is legal but which incites hatred against certain people specified in section 62(3) of the Act is priority content that is harmful to children (PC). Our Protection of Children User-to-user Code includes recommender systems measures regarding PPC and PC. Under PCU E1, certain service providers should (in summary) exclude content indicated potentially to be PPC (including suicide content) from the content feeds of users not determined to be an adult. Under PCU E2, certain service providers should (in summary) exclude or give a low degree of prominence to content indicated potentially to be priority content (including hate) from the content feeds of users not determined to be an adult.

[436] We used the term 'relevant available information' in a similar measure in our Protection of Children User-to-User Code, PCU E1 ('Content recommender systems: excluding potential primary priority content'). We acknowledge that this term is similar to the term 'all relevant information that is reasonably available to a provider', which is a statutory term used in section 192 of the Act to refer to the information that providers should use where making a judgment about whether content is content of a particular kind (for example, making an illegal content judgement). However we have decided to continue to use the term 'relevant available information' for consistency with PCU E1.

14.24 If the service determines that it is not illegal content nor illegal content proxy, this measure would no longer apply to that piece of content and the provider could reinstate the content in users' recommender feeds.[437]

Relevant available information

14.25 We do not propose to be prescriptive about the information that providers should take into account to indicate content is potentially illegal. By 'relevant available information' we mean any type of information or signal that is reasonably available from the operation of the service and that a recommender system can use as an indicator to determine whether content is potentially illegal. All service providers are expected to have existing sources of relevant information. Providers should make use of the relevant information that is available to them, which may vary by service. Depending on how the recommender system is designed and configured, providers could use any of the following as relevant available information: content metadata,[438] content labels or tags,[439] suspicious sources,[440] user feedback[441] (including explicit negative feedback from children under PoC measure PCU E3 or another negative feedback mechanism), detection of content by automated tools,[442] or user reports, including those from trusted flaggers.[443]

14.26 This is a non-exhaustive list of examples based on our understanding of relevant information typically available. Service providers may choose to introduce additional ways to gather relevant information that content may be illegal. Considering a range of signals is likely to be more effective than relying on a single source of information when determining whether there are indicators that content is potentially illegal. Service providers should use a combination of information available them to determine whether content is indicated potentially to be illegal content.

14.27 We expect service providers to configure their systems and processes so that they use relevant available information to exclude content from recommender feeds where there are sufficient indicators that it is illegal content. They should act even if their content

[437] In circumstances where measures PCU E1-3 in the Protection of Children Codes of Practice apply, services should also comply with those measures.

[438] Content metadata is the descriptive information about the content itself which can include various attributes that provide context, structure, and insights into the content, such as type (text, image, video), caption, hashtags, mentions, and engagement (likes, shares, and view counts).

[439] Content labels or tags are visual/textual information to categorise content, these may include moderation labels (such as misinformation, harmful, political), discoverability (such as sponsored, trending), and accessibility (such as trigger warnings).

[440] Suspicious sources are an account or source that has a high number of content reports and exclusion associated with it, the recommender system may treat content from such accounts as harmful or violative.

[441] User feedback is the various types of data that helps the recommender systems learn about users' preferences, behaviour, and interactions with content.

[442] The definition of 'relevant available information' set out in our draft Illegal Content User-to-user Codes includes 'indicators generated by technology used on the service'. This is also the case in our Protection of Children User-to-user Code. [accessed 13 June 2025]. In this consultation, we are proposing to introduce a measure into both Codes which recommends the use of proactive technology to detect content (see Chapter 9). We are also proposing to introduce two further hash matching measures into our Illegal Content User-to-user Codes (see Chapter 11 and Chapter 12). Where services detect content using proactive technology, this is an 'indicator generated by technology used on the service' and should therefore be treated as 'relevant available information' for the purpose of our recommender systems measures.

[443] Trusted flaggers are individuals, non-governmental organisations (NGOs), government agencies, and other entities that have demonstrated accuracy and reliability in flagging content that violates a platform's terms of service. As a result, they often receive special flagging tools such as the ability to bulk flag content.

moderation function has not yet determined the content to be illegal, or where the content may not have entered the content moderation pipeline. This could also include (but is not limited to) content in the pipeline which is suspected to be illegal but is awaiting review. This approach means this measure would offer further protections to users which are complementary to those provided by our various Content Moderation measures.

Effectiveness at addressing risks and benefits

14.28 Given that recommender systems have been identified as a pathway to users being exposed to potentially illegal content, this measure minimises harm by reducing the likelihood of users encountering illegal content on their recommendation feed.

14.29 Existing content moderation measures already support this outcome. However, given the volume of content on user-to-user services and the potential delay in content moderation decisions being reached, it may take time for action to be taken on all illegal content in the content moderation pipeline. Without this measure, there is a risk that recommender systems push a piece of illegal content to significant numbers of users in the interim period between a service provider first having reason to suspect that the content might be illegal, and the point at which it makes an illegal content judgement.

Case study of risk posed by recommender systems

Ofcom analysis and stakeholder engagement in response to the 2024 UK violent disorder following the murder of three girls on 29 July 2024 at a children's summer holiday dance class in Southport, Merseyside,[444] demonstrates the harm that can arise due to recommender systems. This work identified the significant role of recommender systems in the dissemination and amplification of harmful and, in some instances, potentially illegal content which contributed to the real-world violent unrest.

Evidence from stakeholders indicated that potentially illegal content was recommended to users across various social media services, including both hateful content and hateful counter-speech in response.[445] [446] In addition, stakeholders noted the amplification of potentially illegal hate content from accounts with large followings, and some even noted that UK proscribed terror groups were marked as verified and monetised accounts, which saw hate and terror content being actively amplified.[447]

Ofcom's open letter to the Secretary of State for the Department of Science, Innovation, and Technology (DSIT) following our analyses, emphasised the significant role played by recommender systems in disseminating and amplifying hateful and illegal content, which

[444] This incident resulted in the deaths of three children. Nine people were injured in the attack, six of them critically.

[445] Counterspeech can be understood as "any direct response to hateful or harmful speech that seeks to undermine it"; Dangerous Speech Project, n.d. What is Counterspeech? [accessed 23 April 2025].

[446] Hateful counterspeech in this area would be considered as threatening, abusive or insulting words or behaviour used toward another person, in this case someone themselves disseminating hate content, as per our Illegal Content Judgements Guidance (ICJG); Ofcom, Illegal Content Judgements Guidance (ICJG), 2024. [accessed 13 June 2025]

[447] Ofcom / Tech Against Terrorism (TAT), 9 August 2024;[✂;]Ofcom / Institute for Strategic Dialogue (ISD), 13 August 2024;[✂]; [✂]

fed into the instigation of real-world hate and violence and for which individuals were convicted. The letter concluded that "posts about the Southport incident and subsequent events from high-profile accounts reached millions of users, demonstrating the role that virality and algorithmic recommendations can play in driving divisive narratives in a crisis period."[448]

14.30 Unless and until content has been confirmed by a service provider as illegal content under an illegal content judgement (or it is an illegal content proxy), it often remains available on recommender feeds. This means users are able to encounter it, potentially in high volumes, especially if it becomes viral.

14.31 This risk of wider dissemination and/or virality is acute because recommender systems can amplify potentially illegal provocative and/or false content, including as a part of a concerted foreign influence campaign, or in the dissemination of hate. These types of content then typically drive user engagement, even where the nature of such engagement may be negative. Given that illegal content is especially likely to drive user engagement – both positive and negative – it is likely that such content would be increasingly recommended to users where it exists on a service and has not yet been removed through content moderation.

14.32 This measure aims to mitigate this risk by preventing the initial amplification, mitigating for the possible virality of both potential illegal content shared without explicit malicious intent, and potential illegal content shared as part of a concerted foreign influence campaign or as an example of 'trashposting' behaviour.

14.33 This measure aims to reduce the risk of users accidentally or unexpectedly encountering content indicated potentially to be illegal content. It also aims to mitigate the risk that users are recommended more illegal content after seeking it out. This would reduce the risks associated with continuous exposure and ongoing engagement with potentially illegal content. Given the role recommender systems can play in driving the dissemination of illegal and potentially illegal content as set out in paragraphs 14.4-14.11, we provisionally consider that the benefits of the measure would be significant for many users.

14.34 We can identify evidence for the feasibility of this measure in existing industry practices. Many user-to-user services have existing external content that can be excluded from recommender feeds to prevent user exposure to certain kinds of violative, borderline, or low-quality content. This demonstrates that the recommender system can be designed and operated accordingly. Stakeholder responses to our May 2024 Consultation on Protecting Children from Harms Online (May 2024 Consultation) also demonstrate that the measure is feasible.[449]

[448] Ofcom, 2024. Letter from Dame Melanie Dawes to the Secretary of State, 22 October 2024 [accessed 23 April 2025].

[449] Ofcom, Statement: Protecting children from harms online - Ofcom, 2025 [accessed 13 June 2025]

Impacts and Costs

Direct costs of implementation

14.35 In order to exclude content indicated potentially to be illegal content from the recommender feeds of users, providers may incur costs related to implementing the steps we have outlined for the measure (see paragraphs 14.22 of 'Explanation of the measure'). In the following paragraphs we set out our understanding of the activities and costs associated with these steps that may need to be implemented to follow this measure.

14.36 Providers should identify what 'relevant available information' exists: There are different sources of relevant available information that a provider could take into account in identifying potential illegal content. Providers have flexibility in determining how these sources of relevant available information indicate that content is likely to be illegal content (for example, whether they use a model to combine sources of information or not).

14.37 Providers should design the content recommender system to take appropriate account of that 'relevant available information': To implement this measure, service providers should configure their recommender systems to process and take into account any relevant available information about the likelihood of content being illegal. Recommender systems typically process large amounts of information (in the form of signals and data) relating to the safety and integrity of content items to determine whether it should be recommended at all. Based on this, we believe that our proposals in this measure are proportionate, technically feasible, and in conformity with industry practice.

14.38 Providers should ensure the content recommender system operates so that content indicated potentially to be illegal content is excluded from users' recommender feeds unless and until it is determined via content moderation to be legal: service providers should ensure that their recommender systems are configured to action signals about content that is likely to be illegal content. We understand that this could be achieved in a variety of ways. For example, by removing that content completely from the recommendation pool, or by modifying the various ranking algorithms and content filters to exclude content that is likely to be illegal content from the feeds of users. To be clear, we are not proposing that service providers use any specific techniques to secure the intended outcome of this measure. Service providers would likely need to test their recommender system once it has been modified. Additional time may be required to conduct these tests to ensure effectiveness.

14.39 We estimate that undertaking these activities could require approximately six to 18 weeks of software engineering time. We have assumed that this time is matched with an equal amount of time from staff in other non-technical professional occupations (such as lawyers, analysts and product managers). Using our assumptions on labour costs set out in Annex 15, we estimate that one-off direct costs could be in the region of £15,000 to £92,000.[450] [451]

14.40 Depending on how services choose to implement the measure, there could be costs associated with model training such as compute costs – for example, if a model is used to

[450] We have used the same estimate of time to implement this measure as was used in for PoC measure PCU E1. The quantified estimate is higher as we have used the latest wage data released by the Office for National Statistics (ONS). See Annex 15 for a detailed description of our salary assumptions.

[451] Although we have drawn on available evidence and expert input, our quantitative estimates of costs should be interpreted as indicative.

combine relevant available information to assess if content is indicated as potentially being illegal content. There may also be business oversight and coordination costs associated with changing products. While we do not have sufficient information to quantify these costs, we would expect them to largely correlate with the size of the service.

14.41 We would expect service providers to incur ongoing costs to ensure that their solution continues to function as intended. This may involve activities such as observing additional variables in product management to see how the measure is performing. In line with our standard cost assumptions set out in Annex 15, we have estimated this to be 25% of the initial set-up costs, which is approximate to £4,000 to £23,000 per year.

14.42 The existing design of the recommender system would have an impact on the cost of this measure for a given provider. We consider that costs would be higher for services that have more recommender systems operating, and for services that have complex systems (for instance, systems 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 'Benefits of the measure and effectiveness at addressing risks' section, 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 such service providers to implement this measure.

14.43 Services that implement this measure may also be in scope of the similar measures PCU E1 and/or PCU E2 in our Protection of Children Codes which apply to providers of services likely to be accessed by children that have content recommender systems and are medium or high risk for types of content harmful to children.[452] For services implementing one or both Protection of Children measures, the incremental cost of this measure is likely to be lower due to overlapping costs. It is possible that implementation of this measure could be marginally more expensive than Protection of Children measure PCU E1 to ensure and maintain the system's performance given it applies to all users on the service (while the Protection of Children measures can be applied only to the recommender feeds of children).

Indirect costs to services

14.44 This measure may have an indirect cost on service providers because certain business models, including advertising and subscription models, generate revenue for a service in proportion to the number of users and/or their engagement.[453] To the extent that not recommending illegal content to users leads to a loss of revenue for a provider, we consider this entirely justifiable. We recognise that some content which is legal and may have been engaging to users may also not be recommended to users as a result of this measure. However, such content may be reinstated if it is determined to be legal via content moderation processes. In addition, this impact may be limited because we consider that services would typically have a wide variety of other content that engages users which can be recommended to maintain engagement and mitigate any adverse impacts.

14.45 There may also be countervailing benefits of this measure to service providers where this measure decreases users' exposure to illegal content. Some users may increase their usage of services where they do not encounter illegal content, which would be positive for user

[452] Ofcom, Protection of Children User-to-User Code, PCU E1.1 and PCU E2.1 [accessed 13 June 2025]
[453] Ofcom, Illegal Harms Register of Risks, p. 13

engagement and revenues. In addition, advertisers could choose to advertise more on a service where they are confident that their content will not be associated with or positioned close to illegal content within recommender feeds.

Rights assessment

Freedom of expression and association

14.46 As explained in Chapter 3 – Introduction and Approach, Article 10 of the European Convention on Human Rights (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. As with Article 10, Article 11 sets out the right to associate with others. Both are qualified rights, and Ofcom must exercise its duties under the Act in a way that does not restrict these rights unless satisfied that it is necessary and proportionate to do so.

14.47 We recognise that this measure has the potential to interfere with users' rights to freedom of expression and association where content is excluded from recommender feeds. The proposed measure recommends that content is excluded from recommender feeds when 'relevant available information' indicates it potentially to be illegal content. We also recognise that this interference may be greater where a service removes a livestream from recommender feeds in accordance with this measure. Given that livestreams are 'live', removal from recommender feeds may negatively impact engagement with the livestream.

14.48 There is also a risk that some of the content excluded from recommender feeds is subsequently determined not to be illegal content.

14.49 Furthermore, given the types of content the proposed measure relates to, there is a risk that such content relates to:

- the most highly protected forms of speech such as religious expression (which could also impact users' rights to religion or belief under Article 9) or political speech; or
- the kinds of content that the Act seeks to protect, such as content of democratic importance, journalistic content, and content from recognised news publishers.

14.50 In relation to suicide content, we also recognise that the risk of error may lead to the exclusion of content that could help people struggling with mental health issues. On balance, we provisionally consider that the associated impact on freedom of expression would be proportionate given the role the measure would play in reducing the dissemination of illegal content.

14.51 There are also safeguards in place which minimise the risk of interference with the right to freedom of expression. Whilst we propose that content which is indicated potentially to be illegal content should be excluded from recommender feeds unless and until it is determined to be legal, the measure does not recommend exclusion of such content from the service altogether (unless and until it is determined to be illegal content and excluded in accordance with providers' duties under the Act and existing measures). In other words, a user may still be able to encounter the content by searching for it on the service or via a link (where these functionalities are available). Furthermore, the potential impact may be mitigated where content is subsequently reviewed by a provider's content moderation function and is determined not to be illegal. In such circumstances, the content could then be reinstated to users' recommended feeds, which in turn minimises the interference with rights to freedom of expression. We recognise that in the case of a livestreamer, this mitigation would not fully restore the livestreamer to the position they would have been in but for the exclusion, because the livestream would not have been recommended while taking place in real time. However, we consider that there are further mitigations which apply.

14.52 Providers also have incentives to make the process by which content is indicated to potentially be illegal content as accurate as possible, to meet user expectations and to minimise the costs of dealing with complaints.

14.53 Other measures in our Illegal Content User-to-User Codes of Practice provide safeguards against disproportionate interference with freedom of expression. These include a number of the Content Moderation measures (set out in Section C of the Codes). In circumstances where providers are making illegal content judgements,[454] they should consult the Illegal Content Judgements Guidance (ICJG).[455] This provides detailed guidance to assist providers in making illegal content judgements and was prepared with careful regard to rights to freedom of expression. Providers are encouraged to have regard to the ICJG when implementing the measure to assist with correctly identifying when freedom of expression considerations are particularly relevant to certain content. Other measures which provide a safeguard are those in relation to complaints by UK users and affected persons if they consider that the provider is not complying with its duties in relation to freedom of expression or privacy.[456]

14.54 We note that these safeguards are only relevant to the extent that content is reviewed and assessed by a provider’s content moderation function or to the extent that an individual is able to make a relevant complaint which means that it would not apply to every piece of content in scope of this measure.

14.55 We also recognise that more significant impacts to users’ rights to freedom of expression and association could arise if services choose to withdraw the parts of their service using recommender systems, or to withdraw the service from the UK market entirely due to the costs of implementing this measure. However, we have given service providers’ flexibility as to how to implement this measure in a way which minimises the costs of implementation as far as possible. In addition, we consider it unlikely that most services in scope of this measure would withdraw from the UK and expect that many providers would retain incentives to enable users in the UK to continue to user their service.

14.56 We note that engagement with content online is often monetised and the extent to which online content is viewed or shared (via a recommender system) can affect the income of both service providers and content creators who benefit from its amplification. We recognise that excluding any content indicated potentially to be illegal content from recommender feeds could result in lower engagement for user-generated content that is indicated by ‘relevant available information’ to potentially be illegal content but is not in fact illegal content. However, the way service providers have designed their algorithms will predominantly determine the ways in which creators’ content is recommended to other users. As noted in paragraph 14.24, the measure does not require providers to continue to exclude content from recommender feeds if it is assessed to be legal following content

[454] See ICU C1.3(a).

[455] Ofcom, Illegal Content Judgement Guidance, 2024.

[456] See measures ICU D1, D2 and D12 in the Illegal content Codes of Practice for user-to-user services.

moderation. To the extent that this measure may affect the revenue stream for service providers or content creators, we consider it is a proportionate approach to securing compliance with the illegal content safety duties under the Act.

14.57 The measure makes clear that it does not recommend the use of any specific kind of proactive technology, or the use of proactive technology to analyse user-generated content communicated privately or metadata relating to user-generated content communicated privately.[457] This is in line with our approach to PCU E1, the measure in the Protection of Children Code of Practice concerning recommender systems excluding potential primary priority content.[458] We took this approach in response to stakeholder feedback.[459]

14.58 The measure does not recommend that providers use additional technologies to implement the measure. However, we understand that most content recommender systems already rely on technologies (including machine learning or artificial intelligence) or other automated tools to determine which content to recommend to users. Information generated by these technologies and tools would be relevant available information which could be used in the design and operation of the content recommender system to identify content indicated potentially to be illegal content. The potential interference with freedom of expression would therefore turn in part on how accurately the information generated by these technologies and tools can be used to identify content indicated potentially to be illegal.[460] There is limited evidence available on this. However, our view is that many content recommender systems are capable of making very sophisticated recommendations to users (based largely on services' incentives to engage the user). In principle, we consider that the technologies used – together with the flexibility given to providers by the measure's non-prescriptive approach – means that the interference with freedom of expression by this measure is proportionate.

14.59 The implementation of the proposed measure could also have positive impacts on the freedom of expression of users. The removal of content indicated potentially to be illegal content from recommender feeds could result in safer spaces online where users may feel more able to join online communities and receive and impart information which is of particular use to them.

14.60 In these circumstances, and noting (1) that services have discretion about how to design and operate their recommender systems so as to implement the measure’s recommendations and about when content is indicated to potentially be illegal content,

[457] Schedule 4 to the Act, paragraph 13 sets out constraints on our power to recommend measures that describe the use of proactive technology as a way (or one of the ways) of complying with a relevant duty. These include that the measure may not recommend the use of the technology to analyse user-generated content communicated privately, or metadata relating to such content. We consider that it is also appropriate to design measures in accordance with these constraints where a measure does not describe the use of proactive technology but sets out outcomes or other steps which can most readily be secured using proactive technology.

[458] See PCI E1.6.

[459] Ofcom, Volume 4: What should services do to mitigate the risks of online harms to children, p. 433 [accessed 1 May 2025]

[460] Pursuant to paragraph 13 of Schedule 4 to the Act, where one of the ways a service provider could implement a measure recommended for compliance with the illegal content duties is through the use of proactive technology, we (1) may not recommend such technology for content communicated privately; (2) must be satisfied that use of the technology is proportionate to the risk of harm, and (3) must have regard to the degree of accuracy, effectiveness and lack of bias of the technology in question.

and (2) that we do not propose to recommend that content indicated potentially to be illegal content is removed from the service completely until it is determined to be illegal content, we consider that any potential interference to users' rights to freedom of expression is likely to be proportionate. It would also be mitigated by the safeguards set out in paragraph 14.51 and is to be weighed against the serious harm that can occur when illegal content is recommended to users by way of recommender feeds.

Privacy and data protection

14.61 As explained in Chapter 3 – Introduction and Approach, Article 8 of the ECHR confers the right to respect for an individual's private and family life. Any interference with this right must be in accordance with the law, pursue a legitimate aim, be proportionate to the legitimate aim and correspond to a pressing social need. Article 8 underpins data protection legislation with which service providers must comply.

14.62 We consider the privacy and data protection impacts of the proposed measure to be inextricably linked and these are therefore assessed together.

14.63 The measure recommends that service providers take into account all relevant available information to ensure that content indicated potentially to be illegal content is excluded from users' recommender feeds. This may include personal data of users.

14.64 However, the 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 or as part of their content recommender systems. It also does not specify that they should obtain or retain any new personal data in those contexts. We therefore consider that service providers can and should implement the proposed measure in a way that minimises any potential impact on users' rights to privacy. We do not consider that there would be significant additional impacts on users' privacy and data protection rights to those already identified in relation to existing measures on recommender feeds and content moderation.

14.65 Furthermore, providers processing users' personal data would still need to comply with applicable data protection legislation, including in relation to the accuracy of personal data and the right for data subjects to make complaints about how their personal data has been processed and about undue interference with their privacy.

14.66 As far 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 would be used in this way. Service providers should comply with their data protection obligations and refer to the Information Commissioner's Office (ICO) guidance on data protection principles and content moderation.[461] This includes consideration of whether the processing is solely automated (in that it has no meaningful human involvement) and results in decisions that have a legal or similarly significant effect on users.

14.67 Overall, and for the reasons set out in paragraphs 14.61-14.66, we have provisionally concluded that the potential interference with users' rights to privacy is proportionate in light of the benefits from protecting users from relevant harms that it would secure.

[461] Information Commissioner's Office, n.d. Content moderation and data protection | ICO [accessed 21 April 2025]

Which providers should implement this measure

14.68 We propose that this measure would apply to providers of user-to-user services that have a content recommender system and are medium or high risk for at least one of the following kinds of priority illegal content: hate, terrorism, suicide, or foreign interference offence content.

14.69 We have focused the measure on these harms because these are the types of harms where we have sufficient evidence to identify recommender systems as a key risk factor for them, as identified in our Risk Assessment Guidance. For these harms, this measure can create significant benefits by helping to prevent users from encountering illegal content associated with these harms on their recommender feeds. Given these benefits and the strength of the evidence associating recommender systems to these four harm types, we provisionally consider that this measure is proportionate where any one of these risks are identified on a service.

14.70 We do not propose that this measure applies to providers where the risk of the relevant harms is low, because the measure would have limited benefits for safety, if any, while its impacts on service providers and users could be material.

14.71 As set out in the Risk Profiles, services with content recommender systems are likely to have increased risk of hate, terrorism, suicide, or foreign interference offences.[462] Services with content recommender systems are only expected to pose a low risk for the relevant harms if there are significant countervailing factors which reduce their risk levels. For example, because there are comprehensive systems and processes in place to limit exposure of users to this content and evidence shows that they are very effective at minimising likelihood of exposure.

14.72 We consider that it is proportionate to propose that this measure apply to service providers of any size, given our view of the effectiveness of the measure and of the role played by these systems in exposing users to relevant harms. The estimated costs of this measure could be significant, which could have an impact on providers of smaller services and result in a potential loss of user choice if a provider chooses to withdraw their recommender systems or to withdraw the service from the UK market entirely. We provisionally consider that the costs would be proportionate given the materiality of the benefits associated with the measure. This argument is strengthened by the fact that generally we expect costs to scale with the size and complexity of services, meaning that providers of smaller services with fewer resources are likely to incur lower costs than larger services. We have also designed this measure to allow service providers appropriate flexibility in how it is implemented.

Provisional conclusion

14.73 Given the harms this measure seeks to mitigate in respect of illegal content, as well as the risks of cumulative harm that recommender systems pose to users, we provisionally consider this measure appropriate and proportionate to recommend for inclusion in the Illegal Content User-to-User Codes.

[462] Ofcom, Illegal Harms Risk Assessment Guidance and Risk Profiles, 2024. [accessed 1 May 2025]

15. User Sanctions – Introduction

What are user sanctions?

15.1 User sanctions, also known as 'enforcement actions', are actions that a service provider may take against a user who has generated, uploaded or shared illegal content or content harmful to children.

15.2 These actions may include giving the user a warning, suspending or banning the user from accessing the service, or in any other way restricting the user's ability to use the service.

Why does it matter?

15.3 The regulatory framework set up by the Online Safety Act 2023 (the Act) is designed to make life online safer for UK users. User sanctions can mitigate the risks associated with users who spread illegal content and/or content harmful to children by discouraging such behaviour or, in some cases, removing user access altogether.

15.4 Our current User-to-user Codes of Practice (Codes) recommend that, unless not technically feasible, providers should take content moderation action in relation to illegal content and content harmful to children.[463] However, the Codes do not currently include recommendations to take further action against the user who has posted such content. Users may therefore continue generating, uploading and sharing illegal content and content harmful to children. We consider that this presents an ongoing risk of harm to UK users.

15.5 Due to the severity of harm posed by Child Sexual Exploitation and Abuse (CSEA) perpetrators, as well as the pervasive and persistent nature of such offending, our view is that permitting such perpetrators to remain on a service poses a risk of further harm to children.

Interaction with the regulatory framework

15.6 The user sanctions measures that we are now proposing would apply to providers of regulated user-to-user services for the purpose of compliance with their duties in sections 10 and 12 of the Act.[464]

15.7 Under section 10, providers of regulated user-to-user services must take proportionate measures to prevent individuals from encountering priority illegal content by means of the service, mitigate the risk of the service being used for the commission or facilitation of a priority offence, and reduce the risk of harm to users from illegal content.

15.8 Under section 12, providers of regulated user-to-user services likely to be accessed by children must take steps to prevent and protect children from encountering content harmful to children by means of the service, and take or use proportionate measures to

[463] ICU C1, ICU C2, PCU C1 and PCU C2

[464] The duties extend only to the design, operation and use of the service in the UK, and as the design, operation and use affects UK users (section 8(3) of the Act).

effectively mitigate and manage the risk and impact of harm to children from content that is harmful to children.

15.9 The duties set out in the above paragraphs include a requirement to consider appropriate 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 (sections 10(4)(d) and 12(8)(d)).

15.10 In preparing Codes describing measures recommended for the purpose of assisting providers to comply with their duties, Ofcom must consider the principles, objectives and provisions set out in Schedule 4 of the Act.[465]

What this part of the consultation covers

15.11 Chapters 15 and 16 set out details of three proposed new measures concerning user sanctions.

15.12 The first measure is directly targeted at users who generate, upload or share CSEA content (i.e. grooming content and child sexual abuse material content (CSAM)), and those who receive CSAM. We propose that for these users, the most serious sanction – a ban – should be enforced. This means removing access for all accounts associated with the user on the service. We also propose that providers should take steps to prevent the user from returning to the service for the duration of the ban.

15.13 Under the other two measures that we are proposing, providers should prepare and apply a user sanctions policy in respect of users who generate, upload or share illegal content, and content harmful to children that is prohibited on the service. These measures are designed to prevent the future dissemination of illegal content and content harmful to children. These measures do not apply where users generate, upload or share CSEA content, as this content is addressed by the first measure.

15.14 These proposed measures build on measures recommended in the December 2024 Statement on Protecting People from Illegal Harms Online (December 2024 Statement) and the April 2025 Statement on Protecting Children from Harms Online (April 2025 Statement) (ICU C1 and ICU C2, and PCU C1 and PCU C2, respectively).

Related amendments in the Codes of Practice

15.15 We are also proposing related amendments to measures in the Illegal Content User-to user Codes and the Protection of Children User-to-user Code. These amendments are needed to ensure that the user sanctions measures are implemented transparently, effectively and with appropriate safeguards in place for users' rights.

15.16 See Chapter 16 for the explanation of the proposed amendments to the performance target and terms of service measures. See Chapters 15, 16 and 21 for the proposed amendments to the appeals measures.

[465] These include that measures described in a Code may relate only to the design or operation of a service in the UK, or as it affects UK users of the service (Schedule 4, paragraph 11).

16. User banning and preventing return following detection of child sexual exploitation and abuse content

Summary

Not only do CSEA offences cause severe harm, such offending is often pervasive and persistent in nature. Permitting such offenders to remain on a service poses an unacceptably high risk of further harm to children in the UK. CSEA content includes both child sexual abuse material (CSAM) and other online behaviours that amount to a grooming offence as set out in our Illegal Content Judgements Guidance.

Therefore, we propose that users who share, generate or upload CSEA content, and those who receive CSAM, should be banned from the service and prevented from returning.

We are proposing to give providers discretion over the technical means to prevent a user from returning to the service. We are proposing that service providers ban users permanently as a default but recognise there may be exceptional circumstances where a permanent ban is not appropriate. A particularly difficult issue is how to apply this measure to children sharing this type of content. Children can be perpetrators of sexual harm to other children. However, there are circumstances where banning a child may not always be appropriate – for example, when a child is coerced into sharing self-generated CSAM. We are seeking views on how best to address this to mitigate the risk of penalising children who may need particular protection.

Our proposals

Number	Proposed measure	Who should implement this
ICU H3	Providers should ban users who share, generate, or upload CSEA, and those who receive CSAM, and take steps to prevent their return to the service for the duration of the ban	All user-to-user services

Consultation questions

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

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

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

Introduction

16.1 In this chapter, we set out our proposal for a measure that recommends that providers ban users and prevent their return to the service where they have carried out child sexual exploitation and abuse (CSEA) on the service.

16.2 In Chapter 16 of this consultation, we propose a measure recommending that all user-to-user services should prepare and enforce a sanctions policy in respect of users who they are aware have generated, uploaded or shared illegal content and/or illegal content proxy.[466]

16.3 We will outline how we propose this measure works.

Interaction with the regulatory framework

16.4 Regulated user-to-user services must take steps under the Online Safety Act 2023 (the Act) to prevent individuals from encountering priority illegal content, mitigate the risk of the service being used for the commission or facilitation of a priority offence, and reduce the risk of harm to individuals from illegal content.[467] The Act states that providers must take proportionate measures across all areas of a service, including policies on user access, which includes blocking users from accessing the service.[468] Section 8(3) provides that these duties extend only to the design, operation and use of the service in the UK.

16.5 This proposed measure is also relevant to the implementation of existing measures, specifically those related to content moderation. This is because in order to ban a user for CSEA content, a provider will need to have identified that content.[469] This measure also relies on the appeals mechanism in the Illegal Content Codes, which we are proposing to amend (as explained further in paragraphs 15.80-15.86).

Explanation of the harm being addressed

16.6 CSEA encompasses a number of priority offences which we have split into two broad categories for the purposes of assessing risk on services: child sexual abuse material (CSAM) and grooming.[470] CSEA is complex, and instances of abuse and exploitation can involve both CSAM offences and grooming.

16.7 CSAM includes material depicting sexual activity, or indecent or prohibited imagery of children, and can take the form of photographic images and videos.[471] CSAM can be shared online between perpetrators. Children can also generate content that constitutes CSAM, which is known as self-generated indecent imagery (SGII). This may be the result of a child

[466] We recommend that providers of services likely to be accessed by children that prohibit at least one kind of primary priority content (PPC), priority content (PC) and/or non-designated content (NDC) should also have a sanctions policy in relation to users who generate, upload or share content harmful to children and/or harmful content proxy.
[467] Online Safety Act 2023, Section 10(2).
[468] Section 10(4)(d) of the Act.
[469] See Ofcom, 2024, Summary of our decisions for a summary of these measures [accessed 30 March 2025].
[470] Ofcom, 2024, Register of Risks, pp.52-102 [accessed 30 March 2025]. See also Ofcom, 2024, Illegal Content Judgements Guidance [accessed 30 March 2025].
[471] This includes imagery created through generative artificial intelligence (GenAI) tools. It also includes non-photographic material such as drawings, animations, or materials providing advice or encouraging the abuse of a child. Ofcom, 2024, Register of Risks, p.83 [accessed 30 March 2025].

being coerced or manipulated by a perpetrator to send intimate images of themselves. Children may also send or receive this material to/from a similar-aged child as part of a consensual, age-appropriate relationship.[472] While this material may be shared consensually within their relationships, there is a risk that such images will later be circulated without consent. The non-consensual sharing of SGII can have a significant impact on victims, including mental health challenges and negative social consequences. Research by Internet Matters shows that while large numbers of children partake in the sending of SGII, 81% of those aged 13-16 surveyed thought that sharing nude images is always harmful.[473]

16.8 CSAM can include deepfake imagery and content created using generative AI tools. There is growing evidence that both adults and children are using these tools to create increasingly realistic CSAM, including taking existing images of children and using 'nudifying apps' to remove clothing.[474]

16.9 Grooming is the process of building a relationship or emotional connection with a child so that the perpetrator can manipulate, exploit, and abuse them. Grooming is a complicated process that does not follow a set pattern. Sometimes it can take weeks or months, in other cases, it can happen quickly after the first contact. Grooming can involve many stages and is not always limited to sexual conversations, although it can involve coercing or manipulating children into multiple sexual acts.[475]

16.10 Evidence shows that online CSEA is widespread and large-scale. Despite existing content moderation practices, online CSEA is increasing with the Internet Watch Foundation (IWF) observing year on year increases in the number of reports received.[476] The scale of online CSEA presents a significant risk of harm to children. Estimates for online CSEA victimisation suggest that in the last 12 months, one in eight children globally have been victim of the non-consensual taking or sharing of, and/or exposure to, sexual images and video. One in eight children globally have also been subjected unwanted sexualised communication, unwanted sexual questions and/or unwanted requests to perform sexual acts in the last 12 months.[477]

16.11 CSEA can have a significant impact on victims and survivors. This can include long term psychological, emotional, and physical effects, difficulties in forming and maintaining relationships and feelings of guilt, shame, and self-blame that can persist into adulthood. In

[472] The term "self-generated indecent imagery (SGII) is the terminology most commonly used to describe this type of CSAM. For the purposes of this document, we will therefore use this term, however in doing so, we recognise the contentious nature of this language and the complexities surrounding young people's online behaviours.

[473] 84% also stated that social media platforms should do more to prevent young people from sharing such images. Internet Matters, 2023, Online misogyny and image-based abuse research report [accessed 3 February 2025].

[474] In 2023, the BBC reported on more than 20 children in Spain having their fully clothed images manipulated to depict them without their clothes on and then shared online. Hedgecoe, G., 2023. AI-generated naked child images shock Spanish town of Almendralejo - BBC News. The BBC, 24 September [accessed 30 April 2025].

[475] Ofcom, 2024, Register of Risks, p.64 [accessed 30 March 2025].

[476] In 2024 it received 291,273 reports containing CSAM, links to CSAM, or advertised CSAM, a 6% increase compared to 2023 and 14% increase from 2022. Internet Watch Foundation, 2025, IWF 2024 Reports Assessment: Combating Online Child Abuse [accessed 29 April 2025].

[477] Childlight, 2024, Into the Light [accessed 22 April 2025].

the Protect Children Global Survivor Survey 2024, 84% of respondents reported long term consequences stemming from the abuse that they had suffered including depression (55%), anxiety disorders and panic attacks (54%), and physical symptoms such as chronic pain and insomnia (38%).[478]

16.12 Evidence suggests that perpetrators may not initially have a strong motivation to offend. However, over time this becomes entrenched, and research shows "patterns of persistent, conscious engagement with CSAM and other minor-sexualising content".[479] As offending behaviours become increasingly normalised, and without any perceived consequence or sanction, the risk to children increases. Offending becomes more frequent, and perpetrators seek new ways to abuse. An anonymous survey carried out among dark web[480] CSEA perpetrators showed that 37% of respondents sought online contact with a child after viewing CSAM.[481] Similarly, over time, perpetrators are known to coerce victims to sending increasingly more harmful imagery or to engage in increasingly harmful sexual activity including the abuse of other children and/or animals.[482]

Explanation of the measure

16.13 This measure is directly targeted at users who generate, share, or upload CSEA content, and those who receive CSAM.[483] We recommend that for CSEA offences, the most serious sanction – a ban – should be enforced.[484]

16.14 We define a ban as removing existing user access to a service.[485] This means removing access for all accounts associated with the user on the service. We also propose providers should take steps to prevent the user from returning to the service. Consistent with paragraph 11 of Schedule 4 to the Act, to follow this measure providers would need to take

[478] Protect Children, 2024, Global Survivor Survey [accessed 3 February 2025].
[479] A 2021 Meta study into 200 accounts showed perpetrators repeated and were persistent in their abuse of children. Meta (Buckley, J., Andrus, M., Williams, C.), 2021, Understanding the intentions of Child Sexual Abuse Material (CSAM) sharers [accessed 28 April 2025].
[480] The 'dark web' refers to an area of the internet that can only be accessed through particular software. This means networks are encrypted repeatedly, making a user anonymous. It is often accessed for illegal purposes.
[481] Insoll, T, Ovaska, A and Vaaranen-Valkonen, N (2021) CSAM Users in the Dark Web: Protecting Children Through Prevention [accessed 31 January 2025].
[482] IICSA, 2020. The Internet Investigation Report March 2020. [accessed 3 February 2025].
[483] Possession of an indecent photograph of a child (section 160 of the Criminal Justice Act 1988 (CJA 1988)) and section 52A of the Civic Government (Scotland) Act 1982 (CG(S)A 1982)) and possession of a prohibited image of a child (section 62 of the Coroners and Justice Act 2009 (CJA 2009)) are priority offences under Schedule 6 to the Act. While this means receiving CSAM is an offence, it is a defence if the recipient did not see the photo and did not know nor had any reason to suspect it was CSAM, or if they received the image without any prior request and did not keep it for an unreasonable time (section 160(2) of the CJA 1988, section 52A(2) of the CG(S)A 1982 and section 64 of the CJA 2009). Where a provider has reasonable grounds to infer that a defence applies, content is not illegal content for the purposes of the Act (section 192 of the Act). As such, where this chapter refers to users receiving CSAM, it means where a provider does not have reasonable grounds to infer that a defence applies.
[484] This measure would not affect the duty to report to Law Enforcement under section 66 of the OSA.
[485] The term banning is used so to avoid any potential confusion with "user blocking," which is used to describe a feature or function available to users who wish to stop contact with another user (see measure ICU J1 of the Illegal Content Codes of Practice).

these actions in relation to the design and operation of the service in the UK, and as it affects UK users.[486]

16.15 Under section 66 of the OSA, service providers are required to report any UK-linked CSEA to the National Crime Agency (NCA). This measure does not impact on this requirement and providers remain responsible for reporting CSEA even when banning identified users.

Why this measure applies to all types of CSEA

16.16 In the User Access chapter of our December 2024 Statement on Protecting People from Illegal Harms Online (December 2024 Statement), we said we would develop a measure on banning users that have shared material containing child sexual abuse (CSAM). We are now proposing this measure which extends beyond only CSAM, to include all forms of sexual abuse and exploitation as listed within Schedule 6 of the Act.[487] There are several reasons for this:

- Perpetrators often commit multiple CSEA harms against a victim and these behaviours are interconnected.[488] For example, grooming is often used to coerce victims to create self-generated CSAM or to engage in online or offline sexual activity.[489]
- To only ban for certain harms may incentivise perpetrators to shift their offending to circumvent a ban. For example, a ban that only applies to known (as in, previously identified) CSAM may encourage perpetrators to seek novel CSAM from victims.[490]
- Covering all forms of CSEA offers children the greatest level of protection from harm.
- All forms of CSEA can have a significant negative impact on victims and survivors.

Why this measure only applies to CSEA

16.17 We are proposing a user ban for CSEA for three main reasons:

- The severity of this harm;
- The high rate of recidivism; and
- The evidence demonstrating how perpetrators will seek to evade bans to continue their offending and cause more harm. For example, by having multiple accounts and/or setting up new accounts.[491]

16.18 We are also proposing that this measure apply to illegal content proxy where the provider is satisfied that its terms of service prohibit CSEA (which we refer to in this chapter as CSEA content proxy). This is because we recognise that many providers will remove CSEA content in accordance with their terms of service without necessarily making an illegal content

[486] A UK user is either an individual in the UK, or an entity incorporated or formed under the law of any part of the UK (section 227(1) of the Act).

[487] Section 59 and Schedule 6 of the Act.

[488] Section 4 and 5 of Ofcom’s Illegal Content Judgements Guidance (ICJG) set out these offences.

[489] Ofcom, 2024, Register of Risks, p.66 [accessed 30 March 2025].

[490] Known CSAM refers to material that has been detected, as opposed to ‘first-generation’ or ‘novel’ CSAM which may not have been previously shared, re-shared or detected.

[491] Ofcom, 2024, Register of Risks, p.75.

judgement. As such, if this measure did not apply to illegal content proxy, then it could result in CSEA perpetrators remaining on the service.[492]

16.19 Some providers may prohibit nudity content in their terms of service and remove CSAM on that basis. We have considered whether capturing CSEA content proxy in the scope of this measure would be proportionate in that scenario. On such services, users could be banned under this measure for sharing lawful nudity content, which would amount to an interference with their user rights. However, we expect users would be able to appeal against any decision to ban them for posting such content, enabling them to have their access to the service restored and mitigating the impact on their rights. Given this, and the risk of harm that would arise from excluding illegal content proxy from this measure, on balance we consider it proportionate to propose including CSEA content proxy.[493]

16.20 In Chapter 21 of this consultation, we are also proposing that the reporting and complaints measures in the Illegal Content Codes of Practice be amended to enable users to appeal against decisions related to illegal content proxy. As explained further in relation to appeals (see paragraphs 15.80-15.86 and 15.109) we believe this will contribute to the proportionality of this measure.

16.21 Where this chapter refers to 'CSEA' or 'CSEA content', this includes both CSEA content caught by the Act and CSEA content proxy, unless otherwise specified. Where this chapter refers to 'CSAM' or 'CSAM content', this includes both CSAM caught by the Act and CSAM content proxy.

How we propose this measure works

16.22 We propose that users who generate, share, or upload CSEA are removed from the service and prevented from returning. This would apply to all types of CSEA. We also propose that this applies to users who receive CSAM.

16.23 We also expect providers to include banning for CSEA in their terms of service.[494]

Technical implementation

16.24 Banning a user from a service and preventing return would require a provider to take technical steps to achieve this outcome. All user-to-user services should have the capability to remove a user's access from the service. We consider preventing a user from returning to be a necessary part of the measure for it to be effective (see 'Benefits and effectiveness').

[492] For the avoidance of doubt, this measure would apply to users who receive CSAM content proxy except where a provider has reasonable grounds to infer that a relevant defence would apply.

[493] We are aware that a number of large providers do prohibit content related to child sexual abuse, rather than simply prohibiting nudity and removing CSAM on that basis: see for example TikTok, 2024, TikTok's Community Guidelines; X's policy on child safety [accessed 2 March 2025]; and Meta, 2024, Meta's community standard on child sexual exploitation, abuse and nudity [accessed 2 March 2025].

[494] This need not go further than confirming that users will be banned for sharing, generating or uploading CSEA or CSEA content proxy.

16.25 There are multiple different methods providers may use to implement this measure, which vary in costs and effectiveness.[495] We acknowledge that there are limitations to each of these methods and that perpetrators will seek to circumvent a ban. Best practice in this area is likely to include multiple different technologies or indicators. This is because the aim is to increase friction to the point that perpetrators are unable to or are disincentivised from trying to return to the service due to the sheer resources needed to rejoin.

16.26 There are potential complications with banning a user and preventing return on certain services. Many service providers operate multiple services, which raises the question whether a user should be banned from all services operated by the same provider, or just from the service where the CSEA has been shared, generated or uploaded. Some services may operate in a decentralised manner where administration control is distributed. This may lead to parts of the service banning individuals only in localised portions and not the entirety of the service. For services that also offer Single Sign-On (SSO) functionality, any registered users banned from that platform may be prevented from accessing any third-party services previously registered with those SSO credentials.[496]

16.27 There are particularly difficult issues around preventing a user from returning to a service which does not have registered users or user accounts, as it is harder to reliably identify the user. This is because common technical identifiers (such as IP addresses or device characteristics) may be shared by multiple users or changed frequently and in these cases a service provider may not be able to implement verification practices (such as 'Know Your Customer' checks).

16.28 We have considered how to approach these cases. One option would be to exclude service providers without registered accounts from the measure, acknowledging the increased technical difficulty of implementing. However, we do not believe this to be consistent with the aim of this measure which is to protect children by preventing users from being able to commit further CSEA offences. This may also encourage perpetrators to use these types of services to avoid being banned, increasing the risk of CSEA harms being carried out on those services. Given this, and acknowledging the difficulties, we consider it is appropriate to propose this measure applies to providers of services that allow unregistered users.

16.29 For service providers which operate multiple services, we have taken into account the persistent nature of CSEA offending. We therefore provisionally consider it would offer the greatest protection to users and would be proportionate for a user to be banned from any other service operated by the provider, where the provider has reasonable grounds to infer the user has access to those services.

16.30 We have also considered whether to recommend the technical means for preventing a user from returning to the service, or whether to leave this to providers' discretion. Our view is that:

- There is unlikely to be an appropriate 'one size fits all' approach, and our ambition is for this measure to apply as broadly as possible. Methods which are very costly and ineffective for one service provider may be cheap and effective for another. By giving flexibility to providers, we would allow them to implement the measure in a way that is suitable for the size and risk level of their service. This should enable

[495] The two main approaches are robust identity verification for new/existing users (Know Your Customer Checks) and using various technologies to uniquely identify users.

[496] This could also be a positive consequence of banning in some circumstances.

them to achieve the aim of the measure in a cost effective and proportionate way. For example, we consider giving providers flexibility goes some way towards helping services with unregistered users to implement this measure.

- The technology available to achieve this policy's intention is ever-changing. Setting too prescriptive a standard could risk lowering the industry standard or may disincentivise or prohibit providers from implementing new, more effective technology in the future.
- Being prescriptive on how service providers implement this measure could inadvertently alert perpetrators on how to circumvent a ban.

16.31 With these considerations in mind, our provisional view is that providers should have discretion to decide how to implement this measure for their particular service. We are therefore proposing to recommend that service providers take reasonable steps to effectively prevent banned users from returning. What is 'reasonable' will vary by provider, taking into account matters such as the functionalities and size of the service, and its available resources. However, it would ultimately be for Ofcom to decide whether the steps taken by the provider are reasonable.

16.32 We particularly welcome stakeholder feedback on these issues.

Duration of the ban

16.33 Consistent with the aims of this measure, we would expect a ban to reflect the severity of the harm committed, to ensure children are protected by adequately reducing the risk of perpetrators committing further harm, and to serve as a deterrent for other users. We have identified three options regarding the duration of a ban:

- Option 1: Recommend all users found to have shared, generated or uploaded CSEA content, and those who have received CSAM content, are banned and prevented from returning to the service permanently, in all circumstances.
- Option 2: Recommend service providers are given discretion as to what would be an appropriate length of a ban (to reflect the considerations set out in paragraph 15.33).
- Option 3: Recommend that users are banned permanently as a default but allow for exceptional circumstances where a service provider considers a permanent ban would not be proportionate.

16.34 It is important to note that we consider the issue of how to deal with child users separately later in this chapter.

Option 1: Users are banned and prevented from returning permanently

16.35 The primary benefit of this option is that it is the most likely to prevent users from returning to a service and cause further harm, and provides a larger deterrent to users from carrying out CSEA harm. Another benefit of this option is that it reflects what many service providers already do. For example, Reddit implement automated permanent bans for users who share content that has previously been confirmed as CSAM.[497]

16.36 In response to our November 2023 Illegal Harms Consultation (November 2023 Consultation), some stakeholders, such as the Canadian Centre for Child Protection (C3P) and the Marie Collins Foundation, indicated that any account that posts and shares known CSAM should be permanently blocked from the service.[498] The Philippine Survivor Network said that users should be banned permanently due to the likely risk of re-offending.[499] The UK Interactive Entertainment Association (UKIE), a trade body for the video games industry, said it permanently prevents an perpetrator's console from rejoining the console network if they have uploaded CSAM, and that it "[does not] see any reason to do otherwise."[500]

16.37 This position also aligns with the views of other key stakeholders including those with lived experience of harm and is reflective of the lifelong impact that CSEA can have on a victim. Multiple parents participating in a lived experience workshop organised by Ofcom told us that the default should be to ban permanently.[501]

16.38 While ensuring the highest level of protection, recommending a permanent ban would deviate from current criminal sentencing guidelines for these offences. Sentencing for CSEA offences in criminal law is complex and, depending on the offence, sentences can vary from community orders to many years in custody. The courts assess the specific circumstances of each case in order to reach a view on an appropriate sentence. Relevant circumstances can include the severity of the offence (considering culpability and harm), any aggravating factors (such as previous offending) and any mitigating factors (such as having shown remorse).

16.39 Though we have therefore had regard to this, being banned from a service is less of an intrusion of an individual's freedoms than being sent to prison. It is common practice for courts to place long term restrictions on the freedoms of registered child sex perpetrators.[502] In all cases, the restrictions imposed and their duration must be proportionate to the nature of the offending and the risk posed by the individual.

Option 2: Service providers are given full discretion over the length of a ban

16.40 The primary benefit of this option is that a more universal approach (which Options 1 and 3 represent) may not be appropriate for all user-to-user services and for all instances where

[497] Reddit issued 176,679 automated permanent suspensions for users who shared content that had previously been confirmed as CSAM from January 2024 to June 2024. Reddit, 2024, Transparency Report: January to June 2024 - Reddit [accessed 27 March 2025].

[498] C3P response to November 2023 Consultation, p.27; Marie Collins Foundation response to November 2023 Consultation, p.18; Philippine Survivor Network response to November 2023 Consultation, p.17; Segregated Payments Ltd response to November 2023 Consultation, p.14; Snap Inc. response to November 2023 Consultation, p.24. Marie Collins Foundation said that "there should be a zero-tolerance approach for sharing this content due to the significant harm caused to victims". Marie Collins Foundation response to November 2023 Consultation, p.18.

[499] Philippine Survivor Network response to November 2023 Consultation, p.17.

[500] Ukie response to November 2023 Consultation, p.30.

[501] Ofcom/Lived Experience Workshop, 30th April 2025.

[502] For example, those convicted of one of the offences set out in Schedules 3 or 5 to the Sexual Offences Act 2003 – which includes several of the CSEA offences caught by the Act – can be made subject to a Sexual Harm Prevention Order. This can prohibit an individual from engaging in particular activities on the internet. Such an Order must be imposed for at least five years, though can be made for longer or indefinitely until a further Order is made.

CSEA content is shared, generated or uploaded, or where CSAM is received. It may be that providers are best placed to assess the relevant circumstances of the case to reach a view on an appropriate length of ban. In its response to the 2023 Consultation, Snap Inc said that providers are best positioned to make these determinations.[503]

16.41 [✂] suggested that user accounts should be permanently banned in the event of a single instance of malicious sharing of CSAM, but that the penalties for other CSAM violations should take into account the severity of harm and the number of strikes on the user's account. [504]

16.42 A benefit of this option is that it could more closely reflect how sentencing for CSEA is determined in criminal law. However, providers are likely to have less information available to them than a Court when making a sentencing decision, such that it may be difficult for providers to reach an informed view on an appropriate duration for a ban. As noted in paragraph 15.39, a prison sentence is much more severe than a ban from a service and that therefore it is arguably reasonable for the duration to be longer for the latter.

16.43 As noted in paragraph 15.35, many service providers impose permanent bans, at least for known CSAM. Allowing discretion may result in a system where different services offer different durations of ban for the same harm. This may result in some services, which do not impose permanent bans, to be perceived as more favourable to perpetrators, thereby pushing perpetrators to these services and increasing the risk to users on them.

16.44 A further difficulty with this option is how much guidance to give to providers to ensure that any ban duration is sufficient to reflect the severity of the harm, and serves as a sufficient deterrent to ensure users – particularly children – are adequately protected. It may not be appropriate for us to set out factors that providers should consider, in part because making these factors known publicly could be gamed by perpetrators to receive a shorter ban. However, without setting out these factors, we risk both not giving service providers guidance as to how to implement this measure and creating cases where service providers implement bans which do not result in adequate protection for users. Indeed, those with lived experience of online harm have expressed concerns about providers being able to decide the length of a ban in case it is not proportionate to the harm caused by CSEA.[505]

Option 3: Users are banned permanently as a default, but allow for exceptional circumstances

16.45 A benefit with this option, compared with Option 1, is that it would allow for exceptional circumstances where CSEA may be shared without malicious intent. In its response to the 2023 Consultation, Snap Inc. said it generally believed that any user who shares known CSAM should be banned permanently, but that this is often dependent on context, and that any policy in this area should recognise that there are some exceptional circumstances or 'edge cases' where discretion may be appropriate.[506] Though we expect these circumstances to arise only rarely, permanently banning users in such circumstances could amount to a disproportionate interference with their rights, especially if they have no

[503] Snap Inc. response to November 2023 Consultation, pp.24-25.
[504] [✂].
[505] Ofcom/Lived Experience Workshop, 30th April 2025.
[506] Snap Inc. response to November 2023 Consultation, pp.24-25.

recourse to having their ban overturned or its duration shortened. We discuss appeals at paragraphs 15.50-15.52 and 15.80-15.86.

16.46 We have considered in which cases a service provider would apply exceptional circumstances under this option. One possibility is that service providers are given discretion to decide what these exceptional circumstances are. Our current understanding of industry practice (for example via responses to our November 2023 Consultation) is limited to CSAM, so we welcome responses on how service providers could apply this to all types of CSEA. Service providers policies in this area are also rarely made public due to the risk of helping perpetrators bypass the measures.

16.47 The risk with allowing services to apply discretion to the length of a ban, even if only in exceptional circumstances, is that it undermines the harm that CSEA causes to victims and survivors, even where, for example, CSAM is shared unintentionally or without malicious intent (such as in outrage). Suggesting a lesser sanction in these circumstances risks implying that the harm experienced by those in the images is less serious. A concern with specifying a list of exceptional circumstances is that this would provide perpetrators with information which could enable them to circumvent a permanent ban (as noted in paragraph 15.43).

Summary

16.48 Having assessed the options, our provisional view is to recommend users are banned permanently as a default but allow for exceptional circumstances (Option 3). This is because in the vast majority of cases this option would still lead to a permanent ban, and therefore have the benefits of Option 1 (permanent ban in all cases) of preventing users from returning to a service and causing further harm, as well as potentially acting as a stronger deterrent and more closely reflecting the severity of the harm of CSEA than Option 2 (give providers full discretion over the duration of the ban).

16.49 We set out in paragraphs 15.44 and 15.46 the potential risks of being prescriptive about what may constitute an exceptional circumstance, and therefore propose to leave this to service providers to decide. We consider it would be appropriate though to set out the factors a provider should consider when setting the duration of a ban, namely that the duration reflects the severity of the harm, and seeks to protects users (particularly children) by serving as a deterrent against the banned user, and also other users, from carrying out further harm.[507] Our expectation is that circumstances which do not lead to a permanent ban would be extremely limited.

Considering appeals

16.50 Providers have a duty under section 21 to take appropriate action in response to complaints by users who have been suspended or banned, or otherwise had their use of the service restricted as a result of posting illegal content.

[507] We would also include these factors for providers to consider under option 2.

16.51 Under our current reporting and complaints measures at ICU D10, if a service provider determines on appeal that the relevant content is not illegal, the provider should put the user back in the position they were in before the ban was imposed.[508]

16.52 The current measures do not include the option of changing a sanction on appeal where the service provider maintains that the relevant content is illegal. We explore whether to propose amendments to the current reporting and complaints measure to accommodate such appeals later in this chapter (paragraphs 15.80-15.86).

Benefits and effectiveness at addressing risk

16.53 On balance, we consider that banning users for CSEA is likely to be an effective means of combatting this harm, which would deliver significant benefits. Banning users for CSEA is already common practice among some services.[509] Banning can disrupt offending and protect children from future harm by making it harder for perpetrators to access services where they can contact children and commit CSEA offences. In particular, it can prevent and disrupt the following behaviours, as set out in paragraph 15.12:

- The repeated and persistent nature in which perpetrators will abuse children; and
- The escalation of offending behaviours caused by the normalisation of CSEA online.

16.54 There is also a wider benefit of this measure in reducing the likelihood of other users encountering CSAM inadvertently. The vast scale of CSAM online results in members of the public, both adults and children, being inadvertently exposed to this material. Six percent of British adults report having been exposed to CSAM in this way.[510] This can be a traumatic experience and can cause considerable distress.

16.55 Evidence also shows that accidental exposure to CSAM can represent a gateway to offending. An anonymous survey of CSEA perpetrators on the dark web found that over half of respondents reported that they first viewed CSAM accidentally.[511] By banning users who share this material from services, we intend to reduce the amount of CSAM in circulation and therefore reduce the risk of accidental exposure (and the potential creation of new perpetrators).

16.56 We considered whether other sanctions such as strikes and warnings (as set out in Chapter 16 User sanctions policy of this consultation) would be effective actions in some cases, either instead of a ban or before a ban is imposed. In our view, these alternative sanctions would not be effective to sufficiently mitigate the harm of CSEA. We therefore consider banning to be the most effective option to achieve our intended aim of preventing users from being able to commit further CSEA harm.

[508] As a consequence of introducing the defined term ‘age assessment appeal’ (explained in chapter 18 of this consultation), we are also proposing to change the existing defined term ‘appeal’ to ‘content appeal’ in the Illegal Content Codes. This is necessary to distinguish it from an ‘age assessment appeal’.

[509] Evidence from platform transparency reports, such as Reddit, Discord, and TikTok, show how many accounts were banned for CSEA related activity.

[510] Ofcom, 2024, Register of Risks, p.88 [accessed 3 May 2025].

[511] Protect Children (Insoll, T., Ovaska, A. and Vaaranen-Valkonen, N.) 2021, CSAM Users in the Dark Web: Protecting Children Through Prevention [accessed 30 January 2025].

16.57 Notwithstanding the analysis set out in paragraphs 15.53 to 15.56, there are some potential limitations to the effectiveness of a ban. As set out in the technical implementation section (paragraphs 15.24-15.32), the effectiveness of a ban may be limited if perpetrators can circumvent it and continue offending by either creating new accounts or by accessing other accounts they already have on the same service.[512] We have sought to increase the effectiveness of this measure by proposing that users are unable to access the service through any of their accounts, and are prevented from returning to the service.

16.58 We are proposing not to be prescriptive on the technical means a service provider should take to implement this measure and prevent a user from returning to the service. This is partly so that perpetrators are not alerted to ways they could circumvent a ban or return to the service, as well as allowing service providers to have the flexibility to implement the measure in a cost effective and proportionate way. This recognises the difficulty in permanently preventing a user from returning to the service.

16.59 The effectiveness of this measure is dependent on the implementation of existing measures, specifically those related to content moderation, and the appeals mechanism in the Codes.[513] This is because to ban a user for CSEA-related harms, a provider will need to have identified CSEA.

Issues relating to child users

16.60 The issue of how to apply this measure to child users is a complex one. We know that there are multiple circumstances where children may share or receive CSAM online. This may include cases such as:

- child victims and survivors who are coerced into sharing SGII as a result of grooming; and
- children who consensually share or receive SGII as part of an age-appropriate relationship.

16.61 However, children can also display harmful sexual behaviours and can carry out sexual abuse and exploitation, coercing other children to share SGII and to engage in sexual activity. Children can use GenAI and/or nudifying tools to create sexual images of other children for the purpose of bullying, harm, or humiliation.[514] Children can distribute SGII to a wider audience after having received it consensually, to take revenge, harm or humiliate another child. In Snap Inc. research, 54% of 13–15-year-olds who had sent SGII reported that the imagery had spread beyond the intended recipient.[515] As set out in paragraph 15.7, the non-consensual sharing of SGII (referring to images initially sent consensually then sent

[512] This risk was highlighted within Geocomply's response to our November 2023 Consultation: "Despite reporting offending accounts, victims have been known to be re-victimised due to perpetrators' accounts either remaining active or the perpetrator gaining access to the victim through separate accounts when the platform operators blocked, deleted, or removed the initial account. This issue called device recidivism leads to perpetrators being able to re-victimise children online and find new victims on platforms even when they have already been banned or removed by the platform". Geocomply Response to November 2023 Consultation, p.3.
[513] See Ofcom, 2024, Summary of our decisions for a summary of these measures [accessed 25 May 2025].
[514] These images, often referred to as 'nude deepfakes', can have a profound negative impact on victims and is damaging to a child's sense of autonomy and control. Many suffer PTSD, depression and anxiety as a result of these images. Deepfake imagery has already been considered to have attributed to one child suicide. Internet Matters, 2024, Experiences of nude deepfakes research | Internet Matters [accessed 23 May 2025].
[515] Snap Inc., 2023, Digital Wellbeing Index [accessed 3 February 2025]. For those aged 13-17 years, 38% who had sent SGII reported that the imagery had spread beyond the intended recipient.

onwards without consent of the victim) can have a significant impact on victims including mental health challenges and negative social consequences.

16.62 Under the law, sharing, generating or uploading SGII is deemed to be a criminal offence. However, in cases where, following police investigation, there is no evidence of abuse, exploitation or malicious intent, forces can record an 'outcome 21' crime type. This means that a crime has been recorded but it has been deemed not in the public interest to pursue a conviction and as such, no action is taken against the child.[516] As such, in cases involving children, where the child has been coerced into sharing sexual images or it is found that this child is sharing/possessing images exchanged as part of a consensual, age appropriate relationship, outcome 21 is often recorded.

16.63 In applying this measure to children, we want to ensure that children are not being penalised and re-victimised if they have been the victim of abuse or coercion and/or, in the case of older children, penalised when sharing/possessing images exchanged as part of a consensual age-appropriate experience. We have considered three options for addressing this issue in the implementation of this measure:

- Option A: Children are not banned, meaning that children are exempt from the proposed measure.
- Option B: Ban children who commit CSEA harm except in the exceptional circumstances listed in paragraph 15.60 where banning may not be appropriate or proportionate.
- Option C: Ban all children who commit CSEA harm and rely on the appeals process for cases where banning may have a disproportionate impact on the child.

Option A: Children are not banned

16.64 There are some benefits to recommending that services do not ban children:[517]

- Victims and survivors would not be banned. The aim of this proposed measure is to prevent users from being able to commit further CSEA harms, and penalising children who have been coerced into sending SGII is not necessarily consistent with that aim. Banning a victim or survivor would put the blame on the victim and may have a negative impact on them, exacerbating existing feelings of guilt, shame and anxiety associated with their experiences.[518] It may also cut children off from support networks that may help them at a time of vulnerability.
- Banning children could act as a barrier to reporting abuse, as children may choose not to report their experiences for fear that this would result in them being banned.
- This option would not rely on children who have been groomed or coerced going through potentially retraumatising appeals process in order to reinstate their access to a service (see paragraphs 15.50-15.52 and 15.80-15.86 on appeals). Victims may not feel ready to disclose the abuse they have faced or may not recognise themselves as victims which may impact on their ability to appeal.

[516] College of Policing, 2016, Briefing Note: Police Action in response to youth produced sexual imagery (sexting) [accessed 1 May 2025].

[517] These benefits also apply to Option B.

[518] Joleby, M., Lunde, C., Landström, S. and Jonsson, LS., 2020, "All of Me Is Completely Different": Experiences and Consequences Among Victims of Technology-Assisted Child Sexual Abuse, Frontiers in Psychology, 11. [accessed 27 March 2025].

16.65 Furthermore, the sharing of SGII has become an increasingly normal part of older children's relationships.[519] An IWF study into the sharing of nude images among children and young people found that for older participants (aged 13-14 years old and over) felt that the requesting and sharing of nude images was a regular and normalised aspect of daily life.[520] This is supported by research carried out by Snap Inc. that showed that 57% of those aged 13-24 said that either they or a friend had sent or received a sexual image in the previous three months.[521] While we are not endorsing this practice, and there are acute risks for children sharing images even in the context of what may initially be a consensual relationship, the reality is that imposing a ban on children who share SGII in these circumstances may result in children getting unduly banned if the images then get reported. This may have a deterrence effect on children and may help address some of the negative consequences associated with the practice of sharing SGII, but it could arguably be disproportionate to the harm caused (if the images have been shared in a consensual and non-harmful way and are not further distributed). More holistic approaches, including through education, to address the normalisation of risky behaviours may be more appropriate.

16.66 However, there are significant risks with this option:

- Firstly, and importantly, children can sexually abuse other children and may pose a risk to other children online. While there is limited data on the number of children with harmful sexual behaviours, research by the National Society for the Prevention of Cruelty to Children (NSPCC) suggests that around a third of child sexual abuse offences (online and offline) are committed by children and young people.[522] This risk was highlighted by victims and survivors during the lived experience workshop organised by Ofcom, where it was stated that the risk of child-on-child abuse is "very real" and "often overlooked", and participants referred to recent cases involving children being found in possession of over 500 indecent images.[523] To exclude children from the measure would therefore reduce its effectiveness in protecting children from CSEA content.

- Secondly, services could only effectively exclude children from a ban if they know they are children. Highly effective age assurance, as set out in our Part 3 HEAA Guidance, enables services to determine if a user is an adult, and may be used to unlock access to additional content and functionalities for adult users.[524] Users who do not complete an age check should be treated as children by default. Exempting children from the ban under this measure could therefore result in a perverse scenario where adults who have not been verified as adults are treated as children and are not banned for CSEA. This would incentivise adult users not to use highly effective age assurance to prove they are an adult, so that they are treated as a child and not banned under this measure.

[519] It is worth noting here while those under 18 are defined as children under the Act, the age of consent for 'offline' sexual activity in the UK is 16.

[520] IWF and ARU, 2023, Its normal these days. Self-generated child sexual abuse fieldwork findings report. [accessed 1 May 2025].

[521] Snap Inc., 2023, Digital Wellbeing Index [accessed 3 February 2025].

[522] NSPCC, 2024, Statistics Briefing: Harmful Sexual Behaviour [accessed 30 April 2025]. We are also aware that children can be involved in other forms of abuse including acting as part of sadistic interest groups such as the 764 group, where child perpetrators will deliberately target other children to engage in dangerous sexual activity and/or self-harm.

[523] Ofcom/Lived Experience Workshop, 30th April 2025.

[524] Ofcom, 2025, Guidance on highly effective age assurance [accessed 29 April 2025].

16.67 Considering the factors above, our view is that excluding children from the ban is likely to reduce its effectiveness due to the prevalence of children who commit CSEA offences and the perverse incentive on adult users to be classed as children. While there are benefits of this option, these can be achieved by Option B. We are therefore not proposing this option. We will now assess the remaining two options.

Option B: Ban children who commit CSEA harm unless in exceptional circumstances

16.68 Under this option, all users would be banned other than in the exceptional circumstances set out in paragraph 15.60, which would mitigate the risks we have identified for Option A. There are the following benefits with Option B:

- First, some children pose a risk to other children and adult perpetrators can and do conceal their age, and it is right for the ban to cover these cases. This option shares the important benefits listed in paragraphs 15.64-15.65 in relation to Option A.
- Second, it would allow for discretion where providers consider a ban may not be appropriate or proportionate, given the impact on the child user. Such discretion was supported by participants in our lived experience workshop. They stated that, while providers should not allow access to anyone who poses a risk to children, there are some “innocent reasons” why children may share CSEA.[525] Participants suggested that in these circumstances there should be a moderation process allowing for bans to be determined and applied on a “case-by-case basis”.[526]

16.69 As set out in paragraph 15.60, our view is that it may not be appropriate to ban a child in the following two exceptional circumstances:

- Child victims and survivors who have been coerced or manipulated into sending SGII.
- Older children who have sent SGII to another child (or received from another child) as part of a consensual age-appropriate relationship.

16.70 There are potential risks with Option B, particularly around the difficulty and added burden on service providers in identifying whether these exceptional circumstances exist.

16.71 In determining whether these exceptional circumstances exist, providers will need to know whether a particular user is a child. We are not recommending that all user-to-user service providers implement highly effective age assurance if we were to pursue Option B. While service providers that already have highly effective age assurance may wish to use this when implementing this measure, smaller service providers may choose to use other means available to them to determine whether a user is a child as part of the evaluation of whether the exceptional circumstances exist, rather than as a wider age verification policy to be applied to all users.[527]

[525] For instance, cases where an older child has sent SGII (or received SGII from another child) as part of a consensual age-appropriate relationship. One participant in the workshop also disclosed having been banned as a result of their abuse.

[526] Ofcom/Lived Experience Workshop, 30th April 2025.

[527] Other means of determining the age of a user should be based on information that service providers have, which may include contextual information. This should not include information coming exclusively from self-declaration.

16.72 We recognise the difficulties providers would face in reaching a firm view that a user is a child. We nevertheless consider it reasonable to give service providers some discretion about how they deal with child users who share, generate or upload SGII.[528] If they could not satisfy themselves that a user is a child or that an exceptional circumstance applies then the measure recommends that they should ban the user. Indeed, under this option, even if they believed an exceptional circumstance did exist, they could still choose to ban a child user if they considered it appropriate.

Option C: Ban all children who commit CSEA harm

16.73 This option would result in the highest level of protection against CSEA. However, it may be disproportionate and penalise children who have been victims of CSEA.

16.74 We consider there to be three benefits to Option C:

- First, it would mean that children who abuse other children and adult perpetrators who conceal their age would be banned if they shared CSEA content.
- Second, it would remove the burden on providers of determining whether an exceptional circumstance exists, and whether a user is a child.
- Third, it would better protect children from the risks associated with consensual sharing of SGII.

16.75 However, treating all users equally under the measure would result in victims and survivors who have been coerced into sharing SGII being banned causing them further emotional distress and suggesting that they are being punished for the abuse that they have suffered. It could lead to children being banned for a behaviour they did not know to be illegal.[529] It may further deter victims and survivors from reporting their abuse if they are fearful that this would result in them being banned.

16.76 To get their access reinstated, the child would have to be aware of their right to appeal and go through an appeals process. Victims and survivors have told us of the barriers experienced in reporting abuse and how this can particularly affect certain demographic groups such as disabled people.[530] Some children may find the process of having to share information relating to their experiences traumatic. Children may also struggle to recognise the abuse that they have suffered and may experience guilt and shame. This may deter victims and survivors from appealing a ban.[531]

[528] We recognise that the information available to providers will depend on the way in which they identify CSEA content, and as such they may lack the content required to reach a view on whether an exceptional circumstance applies. In those cases, they may need to rely on evidence provided by the user to make an informed choice regarding the implementation of a ban.

[529] Christian Action Research and Education Response to November 2023 Illegal Harms Consultation, p.2: "many children do not appreciate that the sending of illegal material is a risk, and a criminal offence. A 2021 study in Sweden found that half of the teenagers surveyed do not view sending illegal material as a risk, despite knowing that grooming is a possibility. Sending self-generated images has become so normalised amongst many teenagers that they do not know the risks until they become victims".

[530] Ofcom/Lived Experience Workshop, 30th April 2025.

[531] Evidence shows that there are numerous barriers to reporting including: fear of reprisals, concerns around how this may be received, complicated feelings in relation to the abuse they have suffered and/or their abuser and the negative implications on self-representation. Halvorsen et al, 2020, To say it out loud is to kill your own childhood - Halvorsen et al.pdf [accessed 1 May 2025].

Summary

16.77 How to deal with children in the context of banning users for CSEA is a complex and sensitive issue in respect of which there are a range of competing factors that need to be considered. We are therefore seeking feedback from stakeholders on which of the two options would be best to help achieve the overall outcome of protecting children and other users from CSEA in a proportionate way and is not punitive of child victims and survivors. We particularly welcome stakeholder feedback on the exceptional circumstances we have identified, including whether it would be appropriate for providers not to ban children in those circumstances, and whether there are any other circumstances where it may be appropriate not to ban a child for sharing, generating or uploading CSEA content.

16.78 While banning is the focus of this measure, it is not the only action a service provider could consider in response to cases where a child shares, generates, or uploads SGII. In particular, providers should look at how their service design impacts the ease with which children are able to create and share intimate images. Given this content is illegal, services must take steps to remove it. In our user sanctions policy chapter 16 of this consultation, we set out a wider range of actions a provider may take in relation to a user once they are aware that the user has generated, uploaded, or shared relevant content. This includes the provision of information, (for example, about relevant sections of community guidelines and terms of service) or educational prompts to improve users' understanding of relevant topics. These are also wider actions the service should consider applying to children in circumstances when banning may be disproportionate.

16.79 It is important to make clear that this information would be for the purpose of providing support and information, ultimately with the aim of increasing awareness and knowledge and/or reducing these behaviours. Describing these as alternative "sanctions" in this context runs the risk of being victim blaming or shaming, which is not consistent with the aims of this measure.

Considering appeals where the content has been correctly identified as CSEA

16.80 As outlined in 'Considering appeals', our existing reporting and complaints measures only allow for providers to reverse a ban where they have incorrectly identified the relevant content as illegal.[532] This means a provider would be acting inconsistently with our measures if it reversed its decision to ban a child where that child has proved on appeal that one of the exceptional circumstances outlined in paragraph 15.60 applies. This is because even if one of those circumstances applied, the relevant content would still be illegal.

16.81 We have therefore considered the advantages and disadvantages of amending the existing reporting and complaints measures to permit providers to reverse a ban where they have correctly identified content as illegal but, on appeal, they have reasonable grounds to infer that one of the exceptional circumstances apply. In our view, the benefits of maintaining the current reporting and complaints measure (meaning providers should only overturn a ban where they determine that the relevant content is not illegal or illegal content proxy) are as follows:

[532] See ICU D10. Ofcom, 2024, Summary of our decisions, p.4.

- This is consistent with the primary purpose of the measure, which is that users who share, generate or upload CSEA content, and those who receive CSAM,[533] are banned from the service and prevented from returning.
- This would be a stronger deterrent to users from sharing, generating or uploading CSEA content, as bans where CSEA is identified would not be able to be overturned.
- There would be less burden on providers in terms of costs and resources to reinstate a user's access if a ban is overturned.

16.82 The main disadvantage of this option is that appeals are an important rights protection, as banning represents a severe intrusion on a user's freedom of expression and association. In particular, if we allow for exceptional circumstances to be considered by the service when deciding whether to ban a child user in the first instance (option B in paragraph 15.63), there would feasibly be cases where a child may want to appeal for their ban to be overturned where an exceptional circumstance exists but this was not identified at the time of taking the decision to ban.

16.83 Option C (ban all children) would mean any child who is detected as having shared, generated or uploaded SGII in one of the exceptional circumstances identified above is banned from the service without an option to appeal this decision. This could mean a permanent ban for a victim of grooming with no option to appeal, which could be undesirable for the reasons outlined in paragraphs 15.75 to 15.76.[534]

16.84 In light of the risks and disadvantages outlined in paragraphs 15.82-15.83, we consider it appropriate for providers to be able to overturn a ban on appeal, even where the relevant content was SGII, given the potential impact on child users.

16.85 We are therefore proposing to amend the reporting and complaints measures to recommend that an appropriate action providers can take on appeal is to: overturn a ban that was imposed under our proposed measure, where the content was correctly identified as CSEA, but one of the identified exceptional circumstances exists. [535] As set out in Chapter 21 of this consultation, we are proposing to amend the reporting and complaints measures to enable users to appeal against decisions taken in relation to illegal content proxy. Similarly to the current measures, those proposals would recommend providers overturn a ban only where they decide that content was not illegal content proxy.

16.86 We do not expect there to be circumstances where a service provider would overturn a ban for an adult who has shared, uploaded or generated CSEA content, or received CSAM, or for a child who has done this outside of the exceptional circumstances identified. In these cases, we do not recommend providers should be able to overturn a ban on appeal. However, in line with our proposal that a provider may choose not to impose a permanent ban in some exceptional cases, it follows that providers should have discretion to impose a shorter ban on appeal. We are therefore proposing the reporting and complaints measures be amended to reflect this. [536] We would expect these circumstances also to be limited.

[533] As explained at paragraph 15.21, all references in this chapter to 'CSEA content' or 'CSEA', and references to 'CSAM' or 'CSAM content' include references to CSEA content proxy and CSAM content proxy respectively, unless specified otherwise.

[534] There may also be cases where SGII which was initially shared consensually as part of a relationship is reported by a child with malicious intent, in order to get another child banned.

[535] See Annex 7 Addenda to illegal Codes for draft wording.

[536] See Annex 7 Addenda to illegal Codes.

Impacts and Costs

16.87 This section considers the main direct and indirect costs associated with service providers banning users that generate, share or upload CSEA and in preventing these users from returning to the service.

16.88 We expect service providers may incur one-off upfront costs as well as annual ongoing costs, which can vary significantly across services. The proposed measure is scalable and flexible, so costs are likely to depend on a number of factors such as the service size, risk for CSEA, and existing systems and processes in place.

16.89 In this section we consider the potential costs that could be incurred by smaller service providers that are low risk for CSEA. We then consider how these costs could increase for service providers with service size and/or at more risk for CSEA.

16.90 Services within scope of this proposed measure will also be within scope of the user sanction measure (see user sanctions policy Chapter 16 of this consultation) and we expect that there may also be some cost synergies when implementing both measures. For example, we expect that service providers are likely to develop their CSEA banning policy jointly with their user sanctions policy, so policy development costs could be lower than our estimates if accounting for the cost synergies.

Policy development (one-off and ongoing costs)

One-off costs

16.91 We expect service providers to have a policy setting out how to implement this measure and to communicate it with individuals working in content moderation. We are not proposing specific recommendations on how service providers should do this, so they would have full flexibility in determining these aspects.

16.92 We expect a small service provider that has not identified any CSEA in the past will need regulatory and/or legal staff to read, understand and reflect on our proposed measure. We expect this may involve some time from senior management to sign-off on the policy. We estimate these costs to be between £200 to £300.[537] There may also be some additional low cost for communicating the policy to individuals working in content moderation.

16.93 We expect service providers with a higher risk of CSEA would need to set out a more detailed and complex policy. Providers of these services are likely to require more staff time to develop the measure, and therefore may incur higher costs. For example, providers of services that identify CSEA may need to consider in greater detail aspects such as how to assess whether a user account should be removed and the method for removing a user account.

Ongoing costs

16.94 There may be ongoing costs associated with maintaining the relevant policy. These are likely to be costs associated with updating the policy where necessary, and ensuring these updates are communicated to individuals working in content moderation. For a small service provider that has identified no CSEA in the past, we estimate the annual cost could

[537] We assume that this could take 3 hours of a professional occupation staff time and 1 hour of a senior manager's time. We have calculated cost using data on wages released by the ONS for different occupations, as well as our own wage estimate assumptions. See Annex 15 for more details.

be around £50 to £75,[538] with some minor additional costs associated with updating relevant individuals working in content moderation, if needed. However, we generally expect the maintenance costs to increase with the complexity of the policy, so we expect larger service providers or those with a higher risk of CSEA would incur higher maintenance costs.

Changes to system infrastructure/functionality (one-off costs)

16.95 As service providers would have flexibility over the technical implementation of our proposed measure, we expect the upfront costs associated with any changes needed to the system infrastructure/functionality of services to vary significantly, depending on the method used to ban and prevent users from returning.

16.96 We expect a service provider that identifies a relatively low volume of CSEA per year could use their existing user management system to manually remove and prevent user accounts from returning. The user management systems of services can be used to search for a user account based on the relevant user’s email address or username, and subsequently remove a user so that they are unable to access the service and prevented from creating a new account with the same username or email. For a small service using an off-the-shelf content management system (such as WordPress), the functionality to remove and prevent user accounts from returning would already be built into the existing user management systems of services.[539]

16.97 Therefore, we expect that a small service provider that identifies some CSEA per year may choose to mostly rely on their existing systems to remove and prevent users from returning. Therefore, they may be unlikely to incur additional system infrastructure/service functionality costs, although some small service providers could incur some costs depending on how their existing systems work. We also acknowledge that service providers may sometimes incur some additional system infrastructure/functionality costs if implementing our measure for non-registered users/user accounts. However, the proposed measure is flexible to allow services to choose a cost-effective way that minimises the cost burden.

16.98 We expect large service providers and/or providers of services with a high risk of CSEA may need to ban and prevent many users from returning. They may therefore choose to implement more sophisticated methods, including automated bans. While some large services may already have the relevant functionalities required in place, others may not and therefore may incur more upfront system infrastructure/functionality costs.

Applying the ban and preventing return (ongoing costs)

16.99 There may be ongoing costs associated with applying their banning policy, following the identification of CSEA. These are mainly likely to be costs associated with activities such as assessing and removing user accounts identified to have shared CSEA and if needed, determining the length of bans. These activities would most likely require time from individuals working in content moderation.

[538] Consistent with our standard assumption for ongoing costs, we assume that the annual maintenance costs are 25% of the costs of developing the measure. See Annex 15 for more detail on our standard cost assumptions.

[539] wpbeginner, 2025, How to block a WordPress user [accessed 30 April 2025].

16.100 These costs will vary widely between service providers, depending on the amount of CSEA identified and the extent to which the process is automated. We expect that a service provider that identifies no CSEA would not incur such costs. For smaller service providers that identify some CSEA per year, the cost would likely depend on the method they use to ban and prevent users from returning. For example, we expect the costs to be low for a small service provider relying on its existing user management system to manually remove and prevent user accounts. We estimate the cost to be approximately £40 to £80 per account assessed and/or removed, if we assume it takes on average 1.5 hours of a content moderator's time.[540]

16.101 A larger service with a higher risk of CSEA is likely to have to enforce the user banning policy more frequently and therefore its total ongoing costs could be materially higher. However, they are likely to use an automated service to do this, reducing the cost per decision made. We generally expect the costs of applying the measure to increase with the risk of CSEA on a service and therefore to be proportionate to the scale of harm.

Indirect effects

16.102 Our proposed measure could indirectly affect the ongoing costs of implementing some of our existing measures. It could indirectly increase costs if our proposed measure results in more appeals and training for content moderators. On the other hand, where the proposed measure has the effect of discouraging or preventing future CSEA content, the content moderation costs may be reduced. While it is difficult to quantify these indirect effects, we generally expect them to increase with a service's size and risk.

16.103 In addition, the banning of user accounts could lead to a reduction in user numbers and in turn revenues, particularly for services with a high risk of CSEA. We aim to reduce the risk of losing users due to false positives (for example, if the service incorrectly bans users who did not share CSEA content) through the appeals process. However, we do not place significant weight on the loss arising from banning perpetrators. In contrast, our proposed measure could make users feel safer engaging on services or advertisers more attracted to safer services; therefore, potentially increase user numbers and revenues.

Rights assessment

16.104 As explained in Article 10 of the ECHR sets out the right to freedom of expression, Article 8 sets out the right to respect for individuals' private and family life, and Article 11 sets out the right to freedom of association. In essence, restrictions on these rights must be necessary and proportionate – that is, the measure's contribution to its objective must outweigh its adverse impacts and not cause a greater interference than required to achieve the legitimate aim.

16.105 Our assessment of those adverse impacts is therefore to be balanced against the measure's contribution to removing and preventing CSEA content. Parliament has legislated for the CSEA offences to be designated as "priority offences" under the Act, requiring service providers to use systems and processes designed to prevent individuals from encountering content that amounts to one of these offences, to mitigate the risk of the service being used for the commission of a CSEA offence, and to minimise the length of time for which

[540] We note that the actual time required may be higher or lower than our assumption depending on the complexity of the case.

content amounting to a CSEA offence is present. This reflects the very substantial public interest that exists in measures that reduce the prevalence and dissemination of CSEA content online, relating to each of the prevention of crime, the protection of health and morals, and the protection of the rights of others.

16.106 This measure should reduce the levels of CSEA online, by banning perpetrators from a service and therefore reducing the risk of them repeat offending. As well as preventing crime, any reduction in child sexual abuse would protect children's rights not to be subjected to inhuman or degrading treatment under Article 3 of the ECHR, as well as more broadly protecting the fundamental values and essential aspects of private life in relation to children, including their health. The state has positive obligations, owed to children as vulnerable individuals, to reinforce the deterrent effect of criminal law put in place to protect children's rights under Articles 3 and 8 of the ECHR.

Freedom of expression and association

16.107 Banning a user from a service will remove the user's ability to impart and receive information and to associate with others on that service. It represents a significant interference with the user's freedom of expression and association on the service for the duration of the ban. The effect extends to other users who are unable to receive information shared by the relevant user.

16.108 We have explained in paragraphs 15.53-15.59 why we consider the proposed measure to be effective, in particular because of the risk of repeat offending and escalation of harm by users who commit CSEA harms. We considered whether this measure is the least restrictive option to achieve the stated aims and provisionally conclude that it is, given that other options such as strikes or only removing an account (i.e. not removing all accounts and not taking steps to prevent return) would not as effectively reduce the prevalence of CSEA.

16.109 While we therefore consider the proposed measure to be a proportionate interference with user rights where a user has been correctly identified as having generated, shared or uploaded CSEA content, or as having received CSAM, we recognise the risk of false positives (i.e. content being incorrectly identified as CSEA) leading to users being wrongly banned. Where this is the case, users will be able to appeal against the ban and have it overturned if the provider determines that the relevant content was not illegal or illegal content proxy, thereby ending the wrongful interference with their rights.

16.110 We have considered three options for the duration of the ban, each of which will have differing impacts on user rights:

- Option 1 would recommend that users who share, generate or upload CSEA content are banned permanently. Though this would represent a significant interference with that user's rights, the severity of harm caused by CSEA and the risk of repeat offending may mean such interference would be proportionate, as permitting the user to return to the service even after an extended period of time would enable them to resume carrying out harm. Users would be able to appeal against their ban and have it overturned where they did not generate, share or upload CSEA content or receive CSAM, offering a safeguard for the protection of their rights. As a further safeguard, children who share SGII would be able to have their ban overturned in the exceptional circumstances we have identified.
- Option 2 would recommend that providers retain discretion over the length of a ban, to enable them to take into account the specific circumstances in which CSEA was shared,

generated or uploaded, or in which CSAM was received, and reach a view on what would be an appropriate duration. We expect this would enable providers to ensure users' rights are interfered with only to the extent proportionate to the harm committed, and therefore would be the most rights protective option. However, it could reduce the effectiveness of the measure. As with the other options, users would be able to have their ban overturned on appeal.

- Option 3 would recommend that a permanent ban be the default position, but that providers should have discretion to apply a shorter ban in exceptional circumstances. This could be more proportionate, given the possibility of lesser interference with user rights arising from a non-permanent ban, though may make the measure less effective. As with the other options, users would have the opportunity to appeal to have their ban overturned.

16.111 For the reasons set out in paragraphs 16.48-16.49, our preferred option is option 3.

16.112 As explained in paragraph 16.60, we have identified two scenarios where it may not be justified to ban children - where they have shared SGII as a victim of grooming, or where they have sent or received SGII consensually as part of an age-appropriate relationship. As regards the two options we have presented:

- Option B would act as a safeguard for user rights by recommending that providers may not ban children in these exceptional circumstances, which would avoid interfering with their rights. It would therefore contribute to the proportionality of this measure.
- Option C would represent an interference with user rights as children would be banned, alongside adults. We consider it less likely that such an interference would be proportionate given the risks set out in paragraphs 15.75-15.76 of banning children in these circumstances. However, we note that these children will have shared CSAM, which is illegal, or otherwise breached the providers' terms of service. Further, they would be able to appeal to have their ban overturned, bringing an end to the interference. If we were to recommend that providers retain discretion over the duration of a ban, they could choose to impose a shorter ban in these circumstances, leading to a more proportionate interference with user rights.

16.113 Given the severity of the harm of CSEA, the effectiveness of banning, and the safeguards for user rights, we consider the interference from this measure – whichever options we pursue – would be proportionate and justified.

Privacy and data protection

16.114 We expect our proposed measure would give rise to interference with users' right to privacy. The right to respect for private and family life under Article 8 of the ECHR includes the right to establish and develop relationships with others, which would be interfered with on any service on which a user is banned.

16.115 We also expect providers will have to review and process the personal information of users in order to remove them from a service, with an ongoing interference required to prevent them from returning. Providers may also review and process personal information in order to determine whether an exceptional circumstance exists that would justify not banning a child.

16.116 Though the extent of any interference will depend on how a provider implements the measure, we consider that any interference would be proportionate and justified given the severity of the harm caused by CSEA and the effectiveness of our proposed measure. We expect that, where providers interfere to a greater extent with users’ privacy, this should lead to a more effective reduction in harm, for example by more effectively preventing a user from returning to the service, or a greater protection of a user’s other rights, for example by resulting in a child not being banned in one of the identified exceptional circumstances.

16.117 In any case, providers will have to comply with data protection law including minimising data collection and processing.

Which providers should implement this measure

16.118 Implementing this proposed measure across all user-to-user services increases its effectiveness and offers the greatest protection to victims and survivors. This is because evidence shows that perpetrators will move between services, seeking out smaller services perceived to have less scrutiny, to circumvent a ban. [✂][541]

16.119 Implementing the proposed measure across all user-to-user services therefore reduces the risk of this measure inadvertently creating “safe havens” for perpetrators, recognising the potential risk of CSEA on all user-to-user services, including those where this material has not previously been identified.[542]

16.120 The benefits of this proposed measure will be particularly significant in the case of services with high or medium risk of CSEA due to higher volume of CSEA content on such services. We have identified earlier that the costs of the measure are also likely to be higher for such services. However, due to the higher benefits, we consider that it is proportionate for the measure to apply to such services.

16.121 While the benefits of applying the proposed measure to smaller services with a low risk of CSEA will be lower, we nevertheless consider that it is proportionate for such services. There would be clear benefits from reducing the risk of moving to these services to circumvent a ban and evidence suggests that they can be abused (see paragraphs 15.16 and 15.118). We have designed the proposed measure to allow providers the discretion over the technical implementation and the duration of the ban, so they have the flexibility to apply the measure in a cost-effective way that achieves its aim. This helps to minimise costs for smaller services with low risk of CSEA, which we expect to be relatively low (see Cost and impacts, paragraphs 16.87-16.103).

16.122 We therefore propose that it is proportionate to apply this measure to all user-to-user service providers. We are not proposing applying an equivalent measure to search service providers. However, we note such providers are in scope of Measure ICS F2: Provision of child sexual abuse material (CSAM) warnings.

Provisional conclusion

16.123 While we acknowledge the measure is likely to involve both costs to services in implementation, as well as interference with users’ rights to freedom of expression and

[541] [✂]

[542] This was evidenced recently when Mumsnet was targeted with CSAM. Mercer, D., 2025, Mumsnet: Parenting site targeted with child sexual abuse images - BBC News [accessed 1 May 2025].

association and privacy as set out in our Rights assessment section, our provisional view is that the measure would be effective and proportionate.

17. User Sanctions Policy

Summary

Our existing Codes say that providers should swiftly take moderation action against content they determine to be illegal or harmful to children. However, these measures do not prevent the user from posting similar content again. To make online environments safer, we need to seek to address the behaviour of users who choose to post such content.

We are proposing that all user-to-user service providers prepare and apply a sanctions policy in relation to UK users who generate, upload or share illegal content. Providers of services likely to be accessed by children that prohibit content harmful to children should also have a sanctions policy in relation to UK users who generate, upload or share such content.

These policies could help prevent repeat violations, and reduce the spread of such content by educating users, discouraging harmful behaviour or, in some cases, removing user access altogether.

Sanctions can have a significant impact on users, so we propose that providers should consider factors including the severity of the potential harm to UK users, whether the user is a repeat offender, and the potential impact of the sanction on the user being sanctioned when developing their policies.

Providers should prepare a sanctions policy that fits their specific service. Given the wide variety of regulated services, each with different functions, features and levels of risk, we consider it is reasonable to allow for flexibility in how sanctions policies are designed.

Our proposals

Number	Proposed measure	Who should implement this
ICU H2	Providers should prepare and apply a sanctions policy in respect of UK users[543] who generate, upload, or share illegal content and/or illegal content proxy[544], with the objective of preventing future dissemination of illegal content.	All user-to-user services
PCU H2	Providers should prepare and apply a sanctions policy in respect of UK users who generate, upload, or share content harmful to children and/or content that is harmful to children	All user-to-user services, likely to be accessed by children, which prohibit one or more specific kinds of

[543] A UK user is an individual in the UK, or an entity incorporated or formed under the law of any part of the UK (section 227(1) of the Act).

[544] In the Codes, we define "illegal content proxy" as content that a provider determines to be in breach of its terms of service, where: a) the provider had reason to suspect that the content may be illegal content; and b) the provider is satisfied that its terms of service prohibit the type of illegal content which it had reason to suspect existed.

	proxy[545], with the objective of preventing future dissemination of relevant content that is harmful to children.	relevant PPC, PC and/or NDC
ICU C4 (amendment)	Providers should set and record performance targets for their content moderation function covering the time period for taking relevant content moderation action.	Large and/or multi-risk user-to user-services where take down is not technically feasible
PCU C4 (amendment)	Providers should set and record performance targets for their content moderation function covering the time period for taking relevant content moderation action.	Large and/or multi-risk user-to-user services, likely to be accessed by children, where content and/or access level action is not technically feasible

Other amendments

In addition, we are also proposing related amendments to measures in the Illegal User-to-user Content Codes and Protection of Children User-to-user Codes. These amendments are needed to ensure that the user sanctions policy measure is implemented transparently, effectively and with appropriate safeguards in place for users' rights.

See section 'Terms of Service' for the explanation of the proposed amendments to the terms of service measure ICU G1 and PCU G1. See sections ['Considering appeals where the content has been correctly identified as CSEA' in Chapter 16: CSEA User banning, 'Appeals' in this chapter and Chapter 21 'Broadening Appeals' for the proposed amendments to the appeals measures.

Consultation questions

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

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

Definition box: In this chapter, we use the following terms:

Offender: a UK user who generates, uploads or shares content a provider determines to be illegal content/illegal content proxy or content harmful to children/content harmful to children proxy.

Relevant content: illegal content, illegal content proxy, the specific kinds of relevant content harmful to children that are prohibited on a service and content harmful to children proxy.

[545] "Content that is harmful to children proxy" (referred to in this chapter as 'content harmful to children proxy') is content that a provider determines to be in breach of its terms of service where: a) the provider had reason to suspect the content may be relevant primary priority content, relevant priority content or relevant non-designated content; and b) the provider is satisfied that its terms of service prohibit the type of content which it had reason to suspect existed.

What are user sanctions?

17.1 User sanctions, also known as 'enforcement actions', are actions that a service provider may take against a user. They may include giving the user a warning, suspending or banning the user from using the service, or in any other way restricting the user's ability to use the service. In this chapter we make a series of proposals designed to ensure that service providers have appropriate and proportionate policies on user sanctions in place.

What risks do user sanctions address?

17.1 Our current measures recommend that providers should take content moderation action in relation to illegal content, illegal content proxy, content harmful to children where prohibited on the service, and content harmful to children proxy (in this chapter, 'relevant content'). However, the User-to-user Codes of Practice (Codes) do not currently include recommendations to take further action against the user. Users may therefore continue generating, uploading and sharing relevant content. We consider that this presents an ongoing risk of harm to users and could mean there is little incentive for users posting illegal content or content harmful to children to stop doing so.

17.2 With this proposal, our objective is that services have measures in place that reduce the future dissemination of relevant content by influencing user behaviour, including by reducing reoffending.

17.3 In our Illegal Harms Register of Risks (Illegal Harms Register), we identified that for certain kinds of illegal harm (such as incitement to violence and harassment) there is a risk that services do not seek to or are ineffective in preventing banned users from returning to the service after content is taken down.[546]

17.4 In response to the May 2024 Consultation on Protecting Children from Harms Online (May 2024 Consultation), stakeholders suggested that providers should ensure there are clear consequences for users who consistently break the rules – for example, by de-ranking a user's profile or limiting access to their accounts for children if they have shared harmful or dangerous content [547] - and that there should be more severe penalties for distributing inappropriate content.[548]

Interaction with the regulatory framework

17.5 For interaction with the regulatory framework, see Chapter 15.

17.6 When preparing their sanctions policies, providers should also consider the following recommended and proposed measures:

a) ICU H1: Removing accounts linked to proscribed organisations.
b) PCU F4: Signposting children to support when they post content harmful to children.
c) ICU H3 (proposed) : sexual exploitation and abuse banning measure.

[546] The Illegal Harms Register of Risks: Chapter 25: Governance, systems and processes [accessed 13 June 2025]

[547] National Society for the Prevention of Cruelty to Children (NSPCC) response to May 2024 Consultation, pp.34, 47.

[548] Northern Ireland Commissioner for Children and Young People (NICCY) response to May 2024 Consultation, p. 33.

Our proposals

17.7 Having considered stakeholder feedback and relevant evidence, we are proposing to include measures in our Codes to recommend that user-to-user service providers should prepare and apply a sanctions policy in respect of UK users who generate, upload, or share relevant content, with the objective of preventing future dissemination of relevant content.[549]

17.8 We propose two measures which we discuss further in this chapter:

- Measure ICU H2: Providers should prepare and apply a sanctions policy in respect of UK users who generate, upload, or share illegal content and/or illegal content proxy, with the objective of preventing future dissemination of illegal content.
- Measure PCU H2: Providers should prepare and apply a sanctions policy in respect of UK users who generate, upload, or share content harmful to children that is prohibited on the service and/or content harmful to children proxy, with the objective of preventing future dissemination of content harmful to children.

17.9 We propose that Measure ICU H2 should apply to providers of all user-to-user services. We propose that Measure PCU H2 should apply to providers of user-to-user services that are likely to be accessed by children and that prohibit one or more specific kinds of primary priority content (PPC), priority content (PC) and/or non-designated content (NDC). In relation to content that is harmful to children, providers should prepare and apply a user sanction policy for the harms that they prohibit. For example, where a provider only prohibits pornography, its user sanctions policy should cover that harm.

17.10 We do not recommend Measure PCU H2 for service providers who do not prohibit any kinds of content harmful to children. Providers of these services should refer to the Age Assurance and Content Moderation measures in the Protection of Children User-to-user Code.[550]

Options on how services choose to design their sanctions policies

17.11 Given the range of relevant content within scope of the measures, some of which may not be considered illegal or harmful in other jurisdictions, our measures propose that policies should set out sanctions to be imposed on UK users to ensure their impact is proportionate.

17.12 However, we recognise that providers may choose to design their sanctions policies to apply globally. If they wish, providers may incorporate our recommendations within a global policy that also provides for sanctions to be imposed on non-UK users.

[549] Although we are not proposing an equivalent measure for search service providers as search services are not producing the content – and therefore sanctions would not apply – we note such providers are in scope of the provision of child sexual abuse material (CSAM) warnings (Measure ICS F2).

[550] PCU B4, B5, C1 and C2, Protection of Children User to User Code

Measure ICU H2 and PCU H2: Providers should prepare and apply a sanctions policy for UK users generating, uploading or sharing relevant content

How we propose the measures work

17.13 Under these proposed measures, providers should prepare and apply a user sanctions policy in respect of UK users who generate, upload or share relevant content,[551] with the objective of preventing the future dissemination of illegal content and/or content harmful to children.

17.14 These measures do not apply where users generate, upload or share child sexual abuse material (CSEA) and/or CSEA content proxy (for details of the recommendation that applies in these cases, see Chapter 16: CSEA user banning (Measure ICU H3). Similarly, these measures do not apply where a provider has reasonable grounds to infer a user account is operated by or on behalf of proscribed organisation (see Measure ICU H1 which applies in those circumstances).

17.15 We propose Measure ICU H2 should apply to all providers of user-to-user services. We propose measure PCU H2 should apply to all providers of user-to-user services, likely to be accessed by children, that prohibit one or more kinds of PPC, PC and/or NDC. Providers should prepare and apply a user sanctions policy for the harms that they prohibit.

17.16 Policies should set out the sanctions the service provider may apply to users. When setting its policy on the circumstances in which it will apply a sanction, and the seriousness of the appropriate sanction, the provider should have regard to at least the following factors:

- the severity of potential harm to UK users (in the case of PCU H2, child users) if they encounter the relevant content on the service;
- whether the user has previously generated, uploaded or shared relevant content, including whether they were sanctioned for doing so; and
- the potential impact of the type and duration of the sanction on the user being sanctioned.

17.17 When implementing this proposed measure, providers should evaluate the type of policy that is appropriate for their service. Providers have the flexibility to determine the appropriate sanction to apply, having regard to the factors set out in this measure when setting their policy.

Range of sanctions

17.18 There is a range of actions a service provider could take once it becomes aware that a user has generated, uploaded or shared relevant content. Providers can select the appropriate sanction to prevent future dissemination of illegal and/or content harmful to children. For example, for less severe contraventions, providers may wish to consider providing warnings. In response to more severe harms and/or repeat contraventions, providers may decide to restrict or remove a user's access to the service. However, the overall objective of

[551] Where we use the term 'relevant content' we refer to illegal content, illegal content proxy, the specific kinds of relevant content harmful to children that are prohibited on a service and content harmful to children proxy.

the policy should be preventing future dissemination of illegal content and/or content harmful to children.

17.19 In Table 17.1, we set out a non-exhaustive list of sanctions that a provider may impose on a user once they are aware that the user has generated, uploaded, or shared relevant content. These actions may not be mutually exclusive.

Table 17.1: Examples of sanctions and their description

Sanction	Description
Warnings	
Information	Providers may give users information (for example, about relevant sections of community guidelines and terms of service) to ensure they understand that they should not generate, upload or share the relevant content again.
Educational prompts	Providers may give users educational information to improve users' understanding of relevant topics, for example, where a user has shared suicide, self-harm and/or eating disorder content, to ensure they understand that they should not generate, upload or share the relevant content again.
Strike	Providers may consider the use of a strike system (and how long the strike remains on the user's account), a strike being part of a progressive discipline system that may lead to further penalties or sanctions if the number of accumulated strikes reaches a certain threshold.
Account restrictions	
Preventing access to groups and channels	Providers may prevent a user from accessing groups and channels.
Suspend functionality	Providers may restrict a user's access to functionalities. For example, a provider may temporarily prevent a user from creating new groups, sending or forwarding messages, creating new chats with users they have not interacted with before, or uploading new content.
Suspend user	Providers may suspend a user's access to the service for a particular length of time.
Banning (Removing existing user access to service)	Providers may ban a user, removing the user's access to the service.
Preventing user return to the service	Providers may block identifiers associated with a specific account from creating new accounts.

17.20 Our research shows that a range of service providers already use some, and in some cases, all of the sanctions listed in Table 17.1. For example, some providers send their users a warning, while other providers use a strike system to count violations. After receiving a certain number of strikes the user's account may be restricted or disabled depending on the policy that the content violates.[552] [553] [554] Furthermore, providers may take further actions such as temporary and indefinite suspensions or prevent users from creating new accounts. [555] [556] [557]

Factors providers should have regard to when setting their policy

17.21 When setting their sanctions policy, providers should have regard to at least the following factors:

a) the severity of potential harm to UK users (in the case of PCU H2, child users) if they encounter the content identified to be relevant content;
b) whether the user has previously generated, uploaded or shared relevant content, including whether they were sanctioned for doing so; and
c) the potential impact of the type and duration of sanction on the user being sanctioned.

17.22 We consider that having regard to these factors will help providers balance appropriately the need to protect UK users from harms associated with relevant content and users' rights to freedom of expression, association and privacy.

17.23 Providers may choose to take additional factors and context into consideration when setting their sanctions policy.

17.24 A provider may also choose to have regard to the factors when making individual user sanction decisions. For example, the provider of a small, low-risk service may decide that, under its policy, it will consider the factors when relevant content is identified, on a case-by-case basis. We consider that providing this flexibility will allow providers to set and apply a sanctions policy in a way that is appropriate for their service.

Severity of potential harm to UK users if they encounter the relevant content

17.25 We know that many services will consider the severity of harms to users from both illegal content and content harmful to children when setting user sanction policies, and that some harms may be considered more severe and therefore require a stronger sanction.[558]

17.26 When considering the severity of potential harm, we propose that providers should consider the findings of their illegal harms risk assessments and (where applicable) their children's risk assessments. These findings can help providers to carefully consider the

[552] Google response to 2022 Call for Evidence: First phase of online safety regulation, p.29; Meta Platforms Ireland Ltd (Meta) response to 2022 Call for Evidence: First phase of online safety regulation, pp.28-29. Snap Inc. (2025). Snapchat Moderation, Enforcement, and Appeals | Community Guidelines Explainer. [accessed 13 June 2025]

[553] Services may take other information and circumstances into account when taking action. For example, Meta will not implement a strike in certain circumstances, such as when a user has shared their own financial information, or where Meta has additional context about the nature of the violation. Source: Meta Platforms Ireland Ltd. response to 2022 Call for Evidence, pp.28-29.

[554] Reddit. (2025). Content Moderation, Enforcement, and Appeals. [accessed 22 April 2025].

[555] Twitch can remove users from its home page features and marketing campaigns. Twitch. (2025). About Account Enforcements and Chat Bans. [accessed 27 January 2025].

[556] Signal (2025). Signal Terms & Privacy Policy. [accessed 27 January 2025].

[557] Reddit. (2025). Content Moderation, Enforcement, and Appeals [accessed 22 April 2025].

[558] Meta response to 2022 Call for Evidence: First phase of online safety regulation, pp.28-29; Association for UK Interactive Entertainment (Ukie) response to November 2023 Consultation, pp.17-18.

nature and severity of harm, including the number of individuals that could be impacted; the impact on the individuals affected, including how user base demographics affect the way in which users experience harm on the service; and any characteristics of the service that may increase the impact of harm.

Content identified to be illegal content and/or illegal content proxy

17.27 Service providers should have regard to the severity of potential harm to UK users if they were to or may have encountered the illegal content and/or illegal content proxy on the service, including the potential harm to children.

17.28 When considering the severity of potential harm, we propose that providers consider whether the content is priority illegal content. This is because 'severity' is one of the three factors the UK government used to determine its list of priority illegal offences.

Content identified to be content harmful to children and/or content harmful to children proxy

17.29 Providers should have regard to the severity of potential harm to child users if they were to or may have encountered the content harmful to children and/or content harmful to children proxy on the service.

17.30 We propose when considering the severity of potential harm to children, providers 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.

17.31 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 which can be severe and, in some cases, fatal.[559] Even within certain harms, there may be degrees of severity that need to be considered, for example PC such as abuse and hate content.[560]

Whether the user has previously generated, uploaded or shared relevant content, including whether they were sanctioned for doing so

17.32 Service providers should have regard to whether the user has previously generated, uploaded or shared relevant content, including whether they were sanctioned for doing so. Further, all else being equal, we consider that it may be appropriate to sanction repeat breaches more severely than a first breach. Some services may use a strike system to sanction users and track repeat offenders.[561]

17.33 Services may want to consider whether a repeat contravention relates to the same type of content. For example, they may want to consider the appropriate action to take if a user is found to have shared eating disorder content for the third time after a warning was sent the previous two times such content was shared.

17.34 Services may also want to consider the length of time for which previous sanctions are taken into consideration. For example, strikes on TikTok accounts expire after 90 days and

[559] Ofcom, 2025 Children's Register of Risks [accessed 13 June 2025]

[560] Ofcom and the Alan Turing Institute, 2021. Understanding online hate: VSP Regulation and the broader context. [accessed 13 June 2025].

[561] Match Group response to Call for evidence: Second phase of online safety regulation, p.13; Meta (2025). Counting Strikes. [accessed 24 April 2025].; Google response to 2022 Call for Evidence: First phase of online safety regulation, p.29; Snap response to November 2023 Illegal Harms Consultation, p.25.

will no longer be taken into consideration for a permanent account ban.[562] Similarly, Meta does not count strikes on violating content posted over 90 days ago for most violations or over four years ago for severe violations.[563]

17.35 When setting their policies, providers may also consider whether the user was provided with sufficient information as to why they were sanctioned and whether the user had the opportunity to appeal.

Potential impact of the type and duration of the sanction on the user being sanctioned

17.36 Service providers should consider the potential impact of the type and duration of the sanction on the user being sanctioned when setting their policy, and therefore the proportionality of the sanction.[564] We understand that the potential impact on the user may vary depending on the service.

17.37 When setting the type and duration of the sanction, providers may consider how the sanction might impact the user's ability to use the service. For example, users may be able to continue to access content but be unable to generate, upload or share content themselves. Providers may also consider the appropriate duration of sanctions, particularly sanctions that restrict accounts.

17.38 Providers may wish to consider the nature of the relevant harm when considering the potential impact of the sanction on the user and therefore take a nuanced approach. For example, this may include considering the potential vulnerability of the user. Providers may give information and educational prompts to users as part of their user sanctions.[565] Service providers in scope of Measure PCU F4 should signpost children to resources when they post suicide, self-harm, eating disorder or bullying content.

17.39 Providers may also consider whether they have reasonable grounds to believe the user is a child when assessing the potential impact of a sanction. For example, providers may take this into consideration when deciding if banning is the most appropriate sanction. We have considered the options for removing a user's access or preventing their return to the service and the greater impact it may have on child users in Chapter 16: CSEA user banning.

17.40 We are not recommending highly effective age assurance (HEAA) as a part of this measure.

Applying the user sanctions policy

17.41 The user sanctions policy should be applied once a provider has reviewed and assessed the content to be relevant content and has actioned the content, where technically feasible (ICU C1 and C2, and PCU C1 and C2).

17.42 We note that there may be some circumstances where providers consider no further action should be taken once the content has been actioned. We consider it is within the service's discretion as to whether to apply a sanction against a user, and if so, what sanction to apply, in accordance with its sanctions policy.

[562] TikTok. (2025). Content violations and bans. [accessed 25 April 2025]

[563] Meta. Counting Strikes. [accessed 24 April 2025].

[564] We note that providers have a duty under section 22(2) of the Act, when deciding on and implementing safety measures and policies, to have particular regard to the importance of protecting users' right to freedom of expression within the law.

[565] Meta. Suicide, self-injury and eating disorders. [accessed 25 April 2025].

17.43 Where it is not currently technically feasible for a provider to swiftly take action on content it has determined to be relevant content, the provider may consider the benefits of applying a user sanction to prevent the future dissemination of relevant content.[566]

Review of the sanctions policy

17.44 We propose to recommend that providers should regularly review and, where appropriate, update their sanctions policies to ensure they remain fit for purpose. In doing so, providers should take into account the findings of their most recent illegal harms risk assessment and (where applicable) their children's risk assessments. They should also take appropriate account of appeals they have determined during the period covered by the review that related to users being given a warning, suspended, banned, or in any other way restricted from the service.

Appeals

17.45 Providers have a duty to operate a complaints procedure that allows for relevant complaints to be made and provides for appropriate action to be taken by the provider in response to such complaints.[567]

17.46 Our Codes recommend that providers should determine relevant complaints which are appeals.[568] Appeals[569] include complaints by a UK user if the user has been given a warning, suspended, banned, or in any other way restricted from using a service, as a result of content generated, uploaded or shared by the user which the provider considers to be illegal or content harmful to children.[570]

17.47 We propose to amend the measures to recommend that, in determining such appeals, providers should consider whether, in the circumstances of the individual case, the sanction that has been applied is appropriate having regard to its sanctions policy.

17.48 We also propose to add to measure ICU D10 and PCU D10 that if the provider upholds the appeal (in whole or part), the provider should reverse the action taken against the user and determine the appropriate sanction having regard to its sanctions policy.

Terms of service

17.49 We propose to recommend that providers should include information about the sanctions policy in their terms of service to provide transparency to users.[571] This should include at least the types of sanction that the provider may impose under its policy.

[566] For an explanation of taking action on content and technical feasibility, please see our December 2024 Illegal Harms Statement, paragraphs 2.106 to 2.111 and our April 2025 Statement on Protecting Children from Harms Online, paragraphs 11.38 to 11.65. [accessed 13 June 2025]

[567] Providers have a duty under section 21 of the Act to operate a complaints procedure that allows for relevant complaints to be made and provides for appropriate action to be taken by the provider in response to such complaints.

[568] ICU D8 and D9, and PCU D8 and D9.

[569] In the case of content harmful to children, we are referring here to content appeals.

[570] In Chapter 21 we propose amendments to ensure that the appeals measures cover decisions taken on the basis that content was illegal or content harmful to children proxy.

[571] Amendment to measure ICU G1 and PCU G1.

Benefits and effectiveness at addressing risks

17.50 In our Illegal Harms Register, we stated that ineffective procedures around the use of sanctions could increase the risk of harm.[572] [573] [574]

17.51 In response to the May 2024 Protection of Children consultation, the National Society for the Prevention of Cruelty to Children (NSPCC) suggested that providers should ensure that there are clear consequences for users who consistently break the rules.[575] The Northern Ireland Commissioner for Children and Young People (NICCY) Youth Panel said that there should be more severe penalties for distributing inappropriate content.[576]

17.52 Our analysis suggests that well-calibrated sanctions policies can be effective at reducing the dissemination of relevant content online. There is a range of data which supports this view. For example, there is evidence to suggest warnings, including strikes, are effective at reducing reoffending. [577] [578] [579] Evidence suggests that offenders often engage in the same type of conduct online repeatedly, violating the same policy or same feature. In this context, suspending functionalities for users can disrupt offending online.[580] Further, there is evidence that better communication from services regarding the reasons for content removal and sanctions leads to increased perceptions of fairness and reduces likelihood of reoffending.[581] [582] [583] We therefore provisionally consider that this measure would reduce

[572] The Illegal Harms Register of Risk: Section 25 (Governance, systems and processes). p.429. [accessed 13 June 2025]

[573] Thorn, 2021. Responding to online threats: minors' perspectives on disclosing, reporting and blocking. [accessed 21 September 2023].

[574] A survey carried out by Refuge found that 15% of female survivors who had experienced harassment and abuse online said: 'the abuse worsened when they reported the perpetrator or took an action to mitigate the abuse, such as blocking the perpetrator online'. The survey was carried out with 2,264 UK adults, including 1,158 females. 36% of females reported experiencing at least one behaviour suggestive of online abuse or harassment. Source: Unsocial Spaces. [accessed 4 October 2023].

[575] For example, by de-ranking a user's profile or limiting access to their accounts for children if they have shared harmful or dangerous content. Source: NSPCC response to May 2024 Consultation, pp.34, 47.

[576] NICCY response to May 2024 Consultation, p. 33.

[577] A 2021 study found that posting warning messages to users believed to be at risk of suspension can reduce the use of hateful language on some social media networks. Yildirim, M.M. et al. (2023) 'Short of Suspension: How Suspension Warnings Can Reduce Hate Speech on Twitter', Perspectives on Politics, 21(2), pp. 651–663. doi:10.1017/S1537592721002589.

[578] Electronic Arts (EA) state in their transparency report that they have found that warnings are their primary response to violations and that 80% of all violations are a first offence, suggesting that warnings reduce the risk of recidivism. EA Player Safety Transparency Report 2024. [accessed 13 June 2025]

[579] Google said nine in ten people (94%) who receive a first strike never receive a second strike. Google response to 2022 Call for Evidence: First phase of online safety regulation, p.29.

[580] TikTok nine in ten repeat violators on its platform violate the same feature consistently and that 75% violate the same policy category. TikTok. (2023). Supporting creators with an updated account enforcement system. [Accessed 28 January 2025].

[581] Jhaver, S., Appling, D.S., Gilbert, E. and 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*(CSCW), pp.1-33. doi:10.1145/3359294

[582] Suzor, N.P., West, S.M., Quodling, A. and York, J., 2019. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, p.18.

[583] Katsaros, M., Tyler, T., Kim, J. and Meares, T., 2022. Procedural justice and self-governance on Twitter: Unpacking the experience of rule breaking on Twitter. *Journal of Online Trust and Safety*, *1*(3). doi:10.54501/jots.v1i3.38

the amount of illegal content and content harmful to children, users in the UK are exposed to, thus delivering important benefits.

17.53 The range of sanctions that a provider may impose will depend on the service's functionalities and in some cases, there will be limitations to the action that a provider is able to take. We have designed these proposed measures to give service providers a degree of discretion as to how to implement them.

Impacts and costs

17.54 This section considers the direct and indirect impact of the proposed measures on services. We expect costs to vary widely across services. As the proposed measures are scalable and flexible, costs are likely to depend on factors such as the service size, level of risk, the type of sanction, how the service chooses to implement it and existing systems and processes in place.

17.55 To inform our view on the impact on services, we consider the potential costs that could be incurred by smaller service providers that are of low risk. We then consider how these costs could increase with service size and/or risk (for example, the number and/or complexity of relevant harms and total amount of relevant content present on the service).

17.2 Overall, the assessment set out in this section shows that costs are likely to increase with service size and risk, and hence increase with the benefits of reducing the risk of harm:

- We expect the costs for smaller services to be relatively low if they do not identify any relevant content on their service. They mainly relate to developing a simple sanctions policy and familiarising relevant individuals with such policy.
- We expect costs to increase with service size and risks as larger or riskier services need a more complex sanctions policy and associated training. Higher costs also reflect the actions required by the provider to sanction users and consider appeals, following the identification of relevant content.

17.56 All other things being equal, costs will be lower for providers that already have a user sanctions policy in place. For providers which need to include sanctions for both illegal content and content harmful to children in their policies, we expect that the incremental costs would be smaller due to synergies between sanctions for the different types of content. Services within scope of these measures will also be within scope of the CSEA User Banning measure proposed in Chapter 16 and we expect that there may also be some cost synergies with this.

Policy development (one-off and ongoing costs)

17.57 Policy development costs reflect the time required for the relevant staff to develop, write, and approve the user sanctions policy. There may also be costs involved in communicating the detail of this policy to individuals working in content moderation to ensure that they understand how to implement it.

17.58 Given the scalability and flexibility of the proposed measures, we expect these costs to vary widely across services. Providers have significant flexibility over what the policy should look like, for example the length of the policy and the level of detail. The amount of time that will be required to familiarise individuals working in content moderation with the policy will also vary depending upon the complexity of the policy, and the number of staff working in content moderation.

17.59 For example, in the case of a very small, low-risk service, we expect that the sanctions policy could be relatively simple and straightforward.[584] Drafting and agreeing such a policy would require a small amount of time for preparation and sign-off by management. At a minimum, we estimate the costs for a hypothetical very small service that does not identify any relevant risks to be around £200 to £300 for policy development and £50 to £100 per individual working in moderation for communication of the policy.[585] We expect costs for smaller services to increase with the level and complexity of risks identified as they are likely to require a more elaborate sanctions policy.

17.60 Larger services with more functionalities and those with more relevant content may require a more detailed and complex user sanctions policy, which will require more resource to draft, review, and sign off. However, such services are also more likely to already have some form of sanctions policy and training in place, in which case it may be an exercise in ensuring that it is fit for purpose and meets the recommendations of this proposed measure.

17.61 We note that providers may decide to use external or third-party resource to develop their policy. In this scenario there may also be time and labour costs for adapting the policy to the service.

Ongoing policy costs

17.62 There may be ongoing costs associated with maintaining the sanctions policy. This includes periodic review and update of the policy where necessary, and ensuring any changes are communicated to individuals working in content moderation. We expect that at a minimum for a smaller service that does not identify any relevant content and has a very simple policy, the costs involved could be around £50 to £75 per year to review and update the policy and £12 to £25 per year for each individual working in content moderation to communicate such changes.[586] However, we expect larger services or those with higher risk to incur higher costs, as policies are likely to be more complex.

Changes to system infrastructure/functionality (one-off costs)

17.63 Depending on what user sanctions are proposed in the policy and how they are applied, providers may also need to make adjustments to their service infrastructure or functionality to enable them to take actions – such as deletion of groups, restriction of user functionalities, user suspension, user banning, or user blocking – if the service is not currently able to do this. They may wish to develop functionalities that allow them to have a strike system or to suspend functionalities and develop and test educational prompts.

17.64 We expect that larger services are likely to already have the functionality in place to carry out actions such as these, and so additional costs of this are likely to be minimal. Smaller

[584] While such services may be less likely to already have such a policy in place, the associated costs are likely to be lower than for a larger service.

[585] Based on our standard cost assumptions (as set out in Annex 15), if we assume that such a service might require 3 hours of time from professional occupations and 1 hour of time from senior management to develop such a policy, this would represent a cost of £200 to £300. If we assume that for a small service with a relatively simple user sanctions policy, it takes two hours to read and understand the policy, this would represent a cost of £50-100 per individual working in moderation, based on a typical content moderator's salary.

[586] Consistent with our standard assumption for ongoing costs as set out in Annex 15, we assume that the annual maintenance costs are 25% of the costs of developing the policy and providing initial training.

services could potentially face higher costs to do this if they do not currently have the functionality to do so.

17.65 We note that we are not prescribing any specific actions – it is for providers to determine the appropriate sanctions for their service. This means that it is difficult to estimate the costs. For example, for a small low-risk service using an off-the-shelf content management system (CMS) such as WordPress or Drupal, it should be reasonably straightforward for administrators or moderators to send messages to users and prevent them from posting if required. However, the costs of implementing these measures could be higher for some small services, for example those which do not use an off-the-shelf CMS, or which have a high number of unregistered users.

17.66 At the other end of the costs scale, setting up a system to implement a user sanctions policy for a larger and riskier service operating a complex CMS could take a significant amount of time and cost hundreds of thousands of pounds. This is especially true in situations where services use bespoke technology to implement complex features such as automated strike systems, detailed logging, and integration with other features (such as email notifications).

Implementing sanctions (ongoing costs)

17.67 Service providers will also incur costs associated with implementing the sanctions policy. For example, when relevant content is identified on a service, it will be necessary to consider what is the appropriate sanction in accordance with the policy and then apply the sanction.

17.68 These costs will vary widely between services, depending on the amount of relevant content identified, how the sanction is applied and the extent to which the process is automated. However, in general we expect costs to increase with service size and the number and/or complexity of harms on the service. For example, for a small service which is low risk for all types of relevant content; we would expect it to manually apply the policy (such as sending email warnings to the user in question or removing them from the service, as appropriate) only rarely (if at all). We expect that the costs of doing this would be low. If we assume that one hour of content moderator time would be required for each piece of content identified, this would represent a cost of £30 to £50 for each application of the policy.[587]

17.69 A larger, higher-risk service is likely to have to apply the user sanction policy more frequently and may therefore face higher ongoing costs. However, they are likely to consider using an automated service to do this, reducing the cost per decision made. We also consider that the benefits of this measure are likely to scale with the costs, as services which need to apply the user sanctions policy more frequently are likely to be those with a higher amount of relevant content, and therefore higher potential harms.

Indirect effects

17.70 Our proposed measures can indirectly affect the ongoing costs of implementing existing measures. While it is difficult to assess, this impact is likely to increase with service size and risk.

[587] Based on our standard cost assumptions (as set out in Annex 15).

17.71 In dealing with appeals, we expect the costs to be largely covered by our existing measures and requirements for an appeal process under the Act. However, we acknowledge our proposed measures could indirectly result in more appeals, especially for high risk services, resulting in more staff time and higher ongoing costs to implement the existing appeals measures.

17.72 On the other hand, we note that where a user repeatedly breaches a provider's terms of service, the ongoing moderation of such content or behaviour will often increase compliance costs for the service (in terms of the content moderation measures in general, rather than this measure specifically). Where user sanctions have the effect of discouraging or preventing future offending against the service's terms of service, the content moderation costs may be reduced.

17.73 These measures may also increase the cost of providing training to individuals working in content moderation. Such costs are likely to vary depending upon the complexity of the policy, the type of training chosen, and the number of staff to be trained. In general, we expect these costs to increase with service size and the number and/or complexity of harms.

17.74 These proposed measures may also have a potential impact on service user numbers and revenue. Applying sanctions to users could have an impact on user numbers on a service, which in turn could have an indirect financial impact on services, particularly those which offer paid subscriptions. This could be aggravated in the case of false positives (for example, if the service incorrectly sanctions users who did not share relevant content). Conversely, enforcing this measure could make other users feel safer engaging on services, potentially increasing user numbers. However, the impact of this cannot be quantified.

Rights assessment

17.75 This section considers the impact of our proposed measures on users' rights to freedom of expression, to associate with others and to respect for private and family life. As explained in Chapter 15, restrictions on those rights must be necessary and proportionate.

17.76 The measures recommend (a) providers of all user-to-user services prepare and apply a sanctions policy to prevent the future dissemination of illegal content in accordance with their safety duties under section 10 of the Act, and (b) providers of services that prohibit content harmful to children (and therefore do not have HEAA in place) set and apply a sanctions policy to prevent the future dissemination of content harmful to children in accordance with their children's safety duties under section 12 of the Act. Where providers comply with their safety duties by moderating illegal content proxy and/or content harmful to children proxy, we propose that the measures should apply in relation to the generating, uploading or sharing of such proxy content for the purpose of compliance with the duties.

17.77 We consider that the measures pursue legitimate aims. These include the interests of public safety, the protection of health or morals, and the protection of the rights of others (including child users of regulated services). In the case of Measure ICU H2, they also include the interests of national security and the prevention of crime. These legitimate aims correspond to the pressing social need identified by the UK Parliament to make use of internet services safer for individuals in the UK. Providing a safer environment may in turn mean users, including child users, are more able to contribute.

Freedom of expression, association and the right to privacy

17.78 We propose recommending that policies set out sanctions that may be applied to UK users who share relevant content. Sanctions may include removing user accounts or restricting their use. Removing a user's account or restricting their use of it will remove and/or restrict the user's ability to impart and receive information and to associate with others on the service. Depending on the sanction applied, it may represent a significant interference with the user's freedom of expression and association on the service for the duration of the sanction. It may also interfere with the user's right to respect for private and family life, which includes the right to establish and develop relationships with others. The effect extends to other users who are unable to receive information shared by the relevant user. Less serious sanctions such as warnings may have similar effects if they deter users from sharing content because they are uncertain how content moderation and sanctions policy decisions will be applied.

17.79 In the case of Measure PCU H2, the proposal is for users to be sanctioned for sharing content that is harmful to children but not illegal. It has the potential to affect the rights of adult users to share and access legal content and the rights of child users in age groups not judged to be at risk of harm from the relevant types of priority content.

17.80 Where users are sanctioned for content correctly identified as illegal content or content harmful to children, it is more likely that the interference with their rights will be proportionate. There is a risk that users are sanctioned for content that is wrongly identified as illegal or harmful to children. Measure PCU H2 may have a particular impact on the rights of users (including child users) suffering from suicide ideation, self-harm or an eating disorder who may be sanctioned for sharing their personal experience, or who may not share their experience because they are worried about being sanctioned.

17.81 There is also a risk that stronger sanctions are applied because a service does not have the functionality to implement an appropriate range.

17.82 To ensure any interference with users' rights is proportionate, Measure PCU H2 would apply only to services that prohibit content harmful to children. Both measures propose that, in setting its policy on the circumstances in which it will apply a user sanction and the seriousness of the appropriate sanction, a provider should have regard to the severity of harm to UK users (in the case of Measure PCU H2, child users) if they encounter the relevant content, whether the user has previously shared relevant content, and the impact of the sanction on the user who is being sanctioned. Furthermore, providers may consider a range of sanctions including, for example, educational prompts.

17.83 We propose to recommend that providers have regard to the factors listed in paragraph 17.22 when setting their policies. We recognise that the absence of case-by-case consideration of the factors may have an adverse impact on users' rights including the rights of vulnerable users such as those suffering from suicide ideation, self-harm or an eating disorder. A provider may also choose to have regard to the factors when making individual user sanction decisions. However, given the volume of sanctions decisions many providers will need to make, we understand that it is likely to be impracticable for them to apply the factors on a case-by-case basis and that doing so would be likely to result in inconsistent decision-making.

17.84 The process of preparing a policy and considering in advance the appropriate sanction to apply in different circumstances, having regard to the factors set out in the proposed measures, should improve the quality, consistency and predictability of sanctions decisions in a way which may have a positive impact on users' rights. We propose to recommend that providers should regularly review and update their sanctions policies to ensure they remain fit for purpose. Providers should include information about the sanctions policy in their terms of service to provide transparency to users.

17.85 Other measures act as a safeguard for users' rights if they are sanctioned inappropriately, for example those enabling UK users to appeal if their content has been wrongly identified as illegal content or content harmful to children. If, on appeal, a provider reverses a decision that content was illegal content or content harmful to children, it should reverse the action it has taken (ICU D10 and PCU D10). In addition, we are proposing that, in determining appeals relating to a user being given a warning, suspended, banned or in any other way restricted from using a service on the basis that content generated, uploaded or shared by the user is illegal content or content harmful to children (ICU D8 and D9, PCU D8 and D9), a provider should consider whether, in the circumstances of the individual case, the sanction that has been applied is appropriate having regard to its sanctions policy. We also propose to expand ICU D10 and PCU D10 to cover the appropriate action in cases where the provider does not reverse the decision that content was illegal or content harmful to children, and the user appeals the sanction.[588]

17.86 Providers also have reputational and commercial incentives not to sanction users incorrectly or disproportionately.

17.87 Overall, in light of the risk of harm to users posed by illegal content and to child users by content harmful to children, the potential effectiveness of the measures in preventing the future dissemination of relevant content and the safeguards for users' rights, we consider the potential interference of the measure with users' rights to be justified and proportionate.

Privacy and data protection

17.88 Enforcing a sanctions policy will also have implications for users' rights to privacy where a provider reviews content posted by a user (publicly or privately), the user's behaviour, other contextual factors, and user data necessary for a sanction to be applied (including where personal data may need to be retained for longer). The review would only take place where a provider determines that the user has generated, uploaded or shared relevant content, and would be for the purpose of deciding the appropriate action. Personal data may need to be retained for as long as the user sanction remains in force.

17.89 Providers will remain subject to applicable data protection laws, including the principle of data minimisation, in determining how much user information to review. Other measures also act as safeguards for users' privacy rights, in particular complaints processes, in that they promote compliance with the data protection principles of accuracy, fairness and transparency, and assist users to exercise their rights under data protection legislation.

17.90 Overall, we consider the impact on users' rights to privacy is proportionate to the measure's aim of preventing the future dissemination of illegal content and content harmful to children.

[588] In Chapter 21 we propose amendments to ensure that the appeals measures cover decisions taken on the basis that content was illegal or content harmful to children.

Which providers should implement these measures

17.91 Our analysis suggests that applying sanctions to UK users who generate, upload or share relevant content will reduce reoffending and deter users from posting illegal content or content harmful to children. This will in turn reduce the future dissemination of relevant content and harms to users, thereby delivering significant benefits.

17.92 The benefits of these measures will be particularly significant in the case of services with high or medium risk of relevant content due to the higher volume of relevant content on such services. We have identified that the costs of the measures are also likely to be higher for such services. We consider that it is proportionate for the measures to apply to high and medium risk services.

17.93 While the benefits of applying the proposed measures to smaller and lower-risk services will be lower, we still consider that it is proportionate for the measures to apply to such services. There are clear benefits from reducing the risk of future dissemination of relevant content even if there is a relatively low volume of such content identified. Our analysis shows that where the benefits are lower, costs will also be lower, and the providers have the flexibility to apply the measures in a cost-effective way that achieves its aim.

17.94 We are therefore proposing that:

- Measure ICU H2 should apply to all user-to-user service providers; and
- Measure PCU H2 should apply to all user-to-user service providers, likely to be accessed by children, who prohibit at least one specific type of PPC, PC and/or NDC.

17.95 We are not proposing an equivalent measure for search service providers as users of search services are not generating, uploading or sharing content. However, we note that such providers are in scope of Measure ICS F2: Provision of child sexual abuse material (CSAM) warnings.

Provisional conclusion

17.96 Given the harms these measures seek to mitigate in respect of illegal content and content harmful to children, we consider these measures are appropriate and proportionate to recommend for inclusion in the Illegal Harms and Children's Safety Codes.

Measure ICU C4 and PCU C4 (amendment) – Providers should set and record performance targets for their content moderation function

Current measures

17.97 In the Codes, we recommend that providers of large and/or multi-risk services should set and record performance targets covering at least the time period for taking relevant content moderation action and the accuracy of decision making.

17.98 However, where content moderation actions are not currently technically feasible, we did not consider it was reasonable to recommend providers have targets for the time period for taking relevant content moderation action, with the following exceptions:

a) In the December 2024 Statement, we explained that where take down is not currently technically feasible, providers should set targets for the time period for reviewing and assessing at least suspected CSEA and proscribed organisation content, as there are relevant actions that they can take in relation to such content. However, they are not currently recommended to have time-related targets for reviewing and assessing other content.[589]

b) In the April 2025 Statement, we explained where it is not currently technically feasible for providers to take content level action, or it is not technically feasible (or, in the case of PC, proportionate) to apply access controls on content harmful to children, we do not expect providers to set performance targets for the time period for taking relevant content moderation action.[590]

17.99 We note a typographical error in the April 2025 statement which should have stated that we do not expect providers to set time targets where it is not currently technically feasible for them to take content level action on content harmful to children and it is not technically feasible (or, in the case of PC, proportionate) to apply access controls.

Explanation of proposed amendments

17.100 We propose to amend these measures to recommend providers of the following services set performance targets for the time period for reviewing and assessing content, and track whether they are meeting these:

a) services that are large and/or multi-risk for illegal harms, for whom it is currently not technically feasible to take down content; and

b) services that are large and/or multi-risk for content harmful to children and prohibit one or more specific kinds of content harmful to children, for whom it is currently not technically feasible to take content level action on content harmful to children, and it is not technically feasible (or, in the case of PC, proportionate) to apply access controls.

[589] December 2024 Statement. Volume 2 Service Design and User Choice. paragraph 2.201. [accessed 13 June 2025]

[590] April 2025 Statement. Volume 4 What should services do to mitigate the risks of online harms to children. paragraph 11.210. [accessed 13 June 2025]

17.101 In the case of b) in the paragraph above, providers should set performance targets covering the time period for reviewing and assessing the specific kinds of content harmful to children that are prohibited on the service.

17.102 This is because, in the user sanctions policy measures, we propose that such providers can take action against users who post relevant content, even where it is not currently technically feasible to take action on relevant content.

Effectiveness at addressing risks, costs, and rights assessment

17.103 For a full consideration of the effectiveness, costs and rights impact of time-related performance targets, see the December 2024 Statement and the April 2025 Statement.[591]

17.104 Although providers in scope of these amendments (paragraph 17.102) are recommended to set accuracy targets, they are not currently recommended to set time-related targets.[592] We consider striking a balance between timeliness and accuracy is important. The faster such providers review and assess relevant content, the faster they will be able to take appropriate action under their user sanctions policies. We therefore consider these amendments will help reduce harms to UK users.

17.105 We consider that the costs which we consulted on in our November 2023 and May 2024 Consultations and included in our December 2024 and April 2025 Statements are still relevant.[593]

17.106 We note that providers of services in scope of ICU C4 and PCU C4 will already have in place performance targets relating to accuracy. Therefore, the costs outlined in those documents are likely to over-estimate the incremental cost of making this amendment to the measures.

17.107 These proposed amendments to Measures ICU C4 and PCU C4 would form part of a package of measures relating to content moderation for illegal content and content harmful to children. For the reasons explained in our December 2024 Statement and April 2025 Statement, and taking into consideration the benefits to users and other affected persons, we consider that any impact of Measures ICU C4 and PCU C4 on rights to freedom of expression and association, and on privacy and data protection rights, is proportionate.

Provisional conclusion

17.108 We are proposing that providers of the following services should set performance targets for the time period for reviewing and assessing content, and track whether they are meeting these:

a) services that are large and/or multi-risk for illegal harms, for whom it is currently not technically feasible to take down content; and
b) services that are large and/or multi-risk for content harmful to children and prohibit one or more specific kinds of content harmful to children, for whom it is currently not

[591] See section 'Measure on performance targets', pp. 46 to 49 of the December 2024 Statement and section 'Measure PCU C4: Performance targets', pp. 261 to 263 of the April 2025 Statement.

[592] Except, in the case of ICU C4, in relation to reviewing and assessing at least suspected CSEA or proscribed organisation content.

[593] For full details of these cost estimates, please see paragraphs 2.235 to 2.238 of the December 2024 Statement and paragraphs 11.253 to 11.256 of the April 2025 Statement.

technically feasible to take content level action on content harmful to children, and it is not technically feasible (or, in the case of PC, proportionate) to apply access controls.

17.109 Given the harms these measures seek to mitigate in respect of illegal content and content harmful to children, we consider the amended measures appropriate and proportionate to recommend for inclusion in the Illegal Content and Children's Safety Codes.

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

Summary

Highly effective age assurance enables service providers to determine whether or not a particular user is a child. Several of the measures in our Protection of Children Code of Practice for user-to-user services already require providers to implement highly effective age assurance.

In this chapter, we set out the framework needed to use highly effective age assurance to support our Illegal Content User-to-user Codes. We propose that providers should implement it to protect children from grooming and harms associated with livestreaming.

There is a risk that some users could suffer negative impacts as the result of an incorrect assessment of their age – for example, being denied access to certain functionalities. Our guidance about how highly effective age assurance should be implemented helps to reduce this risk. As a further safeguard, we are proposing that service providers allow appeals of highly effective age assurance decisions and take appropriate action where these are upheld.

Our proposals

Number	Proposed measure	Who should implement this
ICU B1	A measure that defines highly effective age assurance for the purposes of the Illegal Content User-to-user Codes and sets out principles that providers should have regard to when implementing an age assurance process.	Services that use highly effective age assurance to determine which UK users of the service are child users for the purpose of targeting measures recommended in the Illegal Content User-to-user Codes at such users.
ICU D15	Providers should allow for appeals of highly effective age assurance decisions and take appropriate action where these are upheld.	Services that use highly effective age assurance to determine which UK users of the service are child users for the purpose of targeting measures recommended in the Illegal Content User-to-user Codes at such users and which are either large or multi-risk.
ICU D16	Providers should allow for appeals of highly effective age assurance decisions and take appropriate action where these are upheld.	Services that use highly effective age assurance to determine which UK users of the service are child users for the purpose of targeting measures recommended in the Illegal Content User-to-user Codes at such users and which are neither large nor multi-risk.

PCU B1	Amendments to codify the definition of highly effective age assurance in the Protection of Children User-to-user Code	A provider in respect of each service likely to be accessed by children that uses highly effective age assurance to determine which UK users of the service are child users for the purpose of targeting measures recommended in the Protection of Children User-to-user Code at such users, their user accounts or their content feeds

Consultation questions

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

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

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

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

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

Background

What is highly effective age assurance?

18.1 Highly effective age assurance refers to a range of processes which enable service providers to determine whether a particular user is a child, and subsequently prevent access to the service or part of the service or target safety measures at children. In our Age Assurance and Children's Access Statement (January 2025 Statement),[594] we confirmed our criteria-based, technology-neutral and future-proofed approach to highly effective age assurance, as reflected in our Guidance on Highly Effective Age Assurance for Part 3 Services (Part 3 HEAA Guidance).[595]

[594] Ofcom, 2025. Age Assurance and Children's Access Statement (January 2025 Statement). [accessed 16 June 2025]

[595] Ofcom, Guidance on Highly Effective Age Assurance for Part 3 Services (Part 3 HEAA Guidance) and also Guidance on highly effective age assurance and other Part 5 duties (Guidance for service providers publishing pornographic content under Part of the Act) [accessed 16 June 2025]

Our proposals for highly effective age assurance in the Illegal Content User-to-user Codes

18.2 Our Illegal Content User-to-user Codes[596] do not currently recommend the use of highly effective age assurance for the purpose of targeting safety measures.[597] However, we are now proposing that services use highly effective age assurance for the purpose of two sets of measures in the Illegal Content User-to-user Codes.

- The first proposal involves the use of highly effective age assurance to target our proposed livestreaming measures at children. In summary, we propose that users should not be able to comment, send gifts, or use reaction functionalities during children's one-to-many livestreams, or be able to record the content. We discuss our proposal and outline our definition of one-to-many livestreams in Chapters 4 and 6.
- The second proposal applies highly effective age assurance to measures protecting children from grooming ('Settings, functionalities and user support', ICU F1 and ICU F2). These measures aim to make it harder for perpetrators to find and interact with child users. We discuss our proposal in Chapter 19.

18.3 For our measures to protect children from grooming, we are recommending that providers either implement highly effective age assurance to apply the measures to users that have not been determined to be an adult, or alternatively apply the measures to all users.

18.4 Our livestreaming measures recommend that providers should implement highly effective age assurance to apply the measures to users that have not been determined to be an adult.

18.5 We set out the detail of these measures in the relevant chapters, including their effectiveness, costs (including the costs associated with the use of highly effective age assurance), impacts and rights assessments (including privacy and data protection).

18.6 To support the use of highly effective age assurance in the Illegal Content User-to-user Codes of Practice, we are proposing some cross cutting measures to ensure it is applied effectively and consistently.

Legal framework

18.7 To the extent that our proposals in this chapter involve amending Codes of Practice, the Online Safety Act 2023 (the Act) sets out that Ofcom must consider a range of objectives, factors and principles when developing such measures.[598] These are set out in detail in Chapter 3 and in the legal framework annex (Annex 17). As part of those considerations, we must ensure that the measures are designed in light of the principles that it is important to

[596] Ofcom, 2025. Illegal content Codes of Practice for User-to-user services [accessed 16 June 2025]

[597] We note that alongside our April 2025 Statement on Protecting Children from Harms Online, we published a consultation titled 'Illegal harms further consultation: User Controls'. Here, we consulted on expanding the application of Measures ICU J1 and ICU J2 in the Illegal Content Codes of Practice, bringing providers of certain smaller user-to-user services that are likely to be accessed by children into scope of these measures where they have relevant risks and functionalities. We proposed that providers should either use highly effective age assurance to offer child users the option to block and mute other users and disable comments on their content or should offer these controls to all users on the parts of the service that are accessible by children. We intend to issue our decision on these proposals as part of the statement that will follow this Additional Safety Measures Consultation.

[598] These are set out in Schedule 4 to the Act.

protect users' and interested persons' rights to freedom of expression and users' rights to privacy. Further, we must incorporate, where appropriate to do so, safeguards for the protection those principles.[599]

Summary of this chapter

18.8 We are proposing to:

a) Define highly effective age assurance in the Illegal Content User-to-user Codes so that it is consistent with other parts of the online safety regime, including the Protection of Children User-to-user Codes. This includes recommending that providers have regard to certain principles when implementing an age assurance process. We are also proposing to change the way in which we present the definition of highly effective age assurance in the Protection of Children User-to-user Code and to mirror this presentation in the Illegal Content User-to-user Codes.

b) Introduce a measure concerning how providers should deal with complaints regarding an incorrect assessment of age based on highly effective age assurance.

c) Make minor amendments to the Part 3 HEAA guidance to broaden its scope to cover highly effective age assurance used in connection with the Illegal Content User-to-user Codes.

18.9 We discuss each of these proposals in the following parts of this section. In general, our proposals align with our approach to highly effective age assurance under the Protection of Children User-to-user Codes (and accompanying statement). Where we propose to take a different approach, we explain the reasons why.

Establishing a consistent approach to highly effective age assurance in the Illegal Content User-to-user Code

Explanation of the measure

18.10 There are three elements to our proposal:

- defining highly effective age assurance;
- introducing a measure which sets out principles that providers should have regard to when implementing an age assurance process; and
- changing the way in which the definition is presented in the Protection of Children User-to-user Code and reflecting this presentation in the Illegal Content User-to-user Codes.

Defining highly effective age assurance in the Illegal Content User-to-user Codes and set out principles that providers should have regard to

18.11 We are proposing to define highly effective age assurance in the Illegal Content User-to-user Codes so that it is consistent with the definition used in the Protection of Children User-to-user Codes and our Part 3 HEAA Guidance. The definition sets out that highly effective age assurance is an age assurance process that fulfils the criteria of technical accuracy, robustness, reliability and fairness.

[599] Schedule 4 to the Act, paragraph 10(1)-(2).

18.12 We are also proposing to introduce a measure into the Illegal Content User-to-user Codes which sets out principles that providers should have regard to when implementing highly effective age assurance for the purpose of other measures which recommend the use of an age assurance process. This will mirror PCU B1 in the Protection of Children User-to-user Codes. It will also provide safeguards to protect users' rights to freedom of expression and privacy (including data protection), which service providers should have regard to when implementing highly effective age assurance.

18.13 The principles are:

a) that age assurance should be easy to use, including by children in the United Kingdom of different ages and with different needs;
b) that age assurance should work effectively for all users, regardless of their characteristics or whether they are members of a certain group;
c) that providers should consider the desirability of ensuring interoperability between different kinds of age assurance
d) that providers should consider the age appropriate design code and the Information Commissioner's Office (ICO) Opinion entitled 'Age Assurance for the Children's Code' published on 18 January 2024.

18.14 The measure will also recommend that providers should ensure users are able to easily access information about what the provider's age assurance process is intended to do and how the provider's age assurance process works. They should be able to do this prior to commencing the age assurance process for the service.

18.15 Lastly, the measure will recommend that the provider should not publish content that directs or encourages UK users to circumvent the highly effective age assurance process (or content controls or access controls used in conjunction with it) on the service.

18.16 While we did not receive feedback on the equivalent measure in the Consultation on Protecting Children from Harms Online (May 2024 Consultation), we received significant feedback on our overall approach to highly effective age assurance, which we addressed in full in our January 2025 Statement and our Statement on Protecting Children from Harms Online[600] (April 2025 Statement).

Codifying the definition of highly effective age assurance in code measures across Protection of Children User-to-user Codes and Illegal Content User-to-user Codes

18.17 We are proposing to change the way in which the definition of highly effective age assurance is presented as compared with the way in which this is currently presented in the Protection of Children User-to-user Code. This will not change the substance, as explained further below.

18.18 Currently, 'highly effective age assurance' is presented as a defined term in the Protection of Children User-to-user Code. Its definition appears at the end of that Code (paragraph 5.10).

18.19 For the Illegal Content User-to-user Codes, we are proposing a different approach. We are proposing to specify in a code measure that a provider should ensure that its age assurance

[600] Ofcom, 2025. Protecting Children from Harms Online [accessed 16 June 2025]

process meets the criteria for it to be considered highly effective. We discuss this further in the 'Impact of the proposals' section below.

18.20 We are also proposing to amend the Protection of Children User-to-user Code to adopt the same approach. In essence, we are proposing to move the elements of highly effective age assurance from the definition into PCU B1, which currently sets out the matters which providers must have regard to when implementing an age assurance process. This approach ensures consistency and provides clarity to service providers, and we discuss the impact further below.

Impacts and Costs

18.21 We assess the impact of highly effective age assurance as used in connection with ICU F1 and F2 and livestreaming measures in Chapter 19, and Chapter 6 respectively. This section focusses on the impact of the measure set out in this chapter, which relates to the definition of highly effective age assurance.

18.22 The definition will apply to service providers that implement an age assurance process for the purpose of acting in accordance with any of the recommendations across the Illegal Content User-to-user Codes. It will therefore assist providers in implementing other measures. Defining highly effective age assurance in the Illegal Content User-to-User Codes also gives legal weight to the definition (in contrast to the non-statutory Part 3 HEAA guidance) which is intended to assist providers but does not have the same legal effect.

18.23 The proposed measure sets out principles that providers should have regard to when implementing highly effective age assurance. This will have a positive impact on users undergoing highly effective age assurance. For example, the measure will ensure that highly effective age assurance is easy to use, works effectively for all users, and that providers do not publish content which directs or encourages its circumvention.

18.24 Introducing this measure will also ensure alignment with the Protection of Children User-to-user Code. By having a consistent definition of highly effective age assurance across both User-to-user Codes, Part 3 services that are in scope of relevant measures in both codes will be able to implement a consistent highly effective age assurance process, reducing cost and complexity. It also provides clarity and certainty to providers as to the circumstances in which they will benefit from the 'safe harbour' provided by Codes of Practice. It may also assist Ofcom in taking enforcement action where a provider does not comply with the measure.

18.25 The proposed measure will also help to safeguard users' rights to freedom of expression and privacy (including data protection). For example, the measure recommends that providers have regard to the ICO age-appropriate design code[601] and ICO Opinion entitled 'Age Assurance for the Children's Code'[602]. It also recommends that providers have regard to the principle that age assurance should work effectively for all users regardless of their characteristics. This would limit the risk that some adults may find it more difficult to assure their age under certain methods (which consequently could impact their freedom of expression). We are also proposing to recommend that providers allow age assessment appeals, as discussed below, and this will also safeguard users' rights.

[601] ICO Age appropriate design: a code of practice for online services [accessed 28 May 2025]

[602] ICO Age assurance for the Children's code [accessed 28 May 2025]

18.26 The measure also codifies the existing definition of highly effective age assurance (in the Protection of Children User-to-user Code) in code measures. This change is intended to assist providers by setting out the elements of the highly effective age assurance definition in one place in both Codes. Adopting this approach means that, when we are identifying measures which act as safeguards for human rights, we can specifically refer to the elements of the highly effective age assurance definition as a safeguard. We consider that this will have a positive impact on users' rights. It is also designed to ensure consistency across our wider Codes and guidance. We are of the view that by making this change the intended effect of the four criteria at paragraph 18.13 forming part of that definition will be clearer.

Which providers should implement the measure

18.27 To the extent that we are proposing to introduce ICU B1 into the Illegal Content User-to-user Codes, we propose that it will apply to user-to-user services that use highly effective age assurance to determine which UK users of the service are child users to target illegal harms measures at children. To the extent that we are proposing to amend PCU B1, this currently applies to providers likely to be accessed by children that use highly effective age assurance to identify which UK users are child users to target protection of children measures at children. We are not proposing to change this, other than a minor change to replace the word 'identify' with 'determine'. This is because 'determine' more accurately describes our policy intent and ensures consistency with other language used in the Protection of Children User-to-user Code[603] e.g. 'determined to be an adult through the use of highly effective age assurance'.

Proposed measures: appropriate action for age assessment appeals

Introduction

18.28 While we consider that the use of highly effective age assurance will lead to the effective application of safety measures for children, the age of some users may be incorrectly assessed. This may have an impact on the online experience and rights of a user whose age is incorrectly assessed, as well as on other users.

18.29 For example, under our proposed livestreaming measures set out in Chapter 6 on livestreaming, we are recommending that users should not be able to comment, send gifts or use reaction functionalities during children's one-to-many livestreams, or be able to record the content. In this context, the incorrect assessment of an adult user's age, resulting in them being considered a child, would have an impact on their ability to receive certain kinds of communication from other users while livestreaming. Similarly, we are proposing that providers may use highly effective age assurance to target existing measures protecting children from grooming ('Settings, functionalities and user support', ICU F1 and ICU F2) at children. Under these proposals, the incorrect assessment of the user's age could lead to some of their settings being off by default and to them receiving certain supportive information. We explain these proposals in greater detail in the relevant Chapter.

[603] Ofcom, 2025. Protection of Children Code of Practice for User-to-user services [accessed 16 June 2025]

18.30 We are therefore proposing to recommend that service providers allow appeals where a user considers their age has been incorrectly assessed and take appropriate action where those appeals are upheld. We are proposing two measures in this regard: one which applies to large or multi-risk services using highly effective age assurance and another which applies to other services (that are neither large nor multi-risk) using highly effective age assurance. These proposed measures are similar to PCU D11 and PCU D12 in the Protection of Children User-to-user Codes ('Appropriate action for age assessment appeals (services that are large or multi-risk)' and 'Appropriate action for age assessment appeals (services that are neither large general search services nor multi-risk') respectively), although there are some differences in the following paragraphs.

Explanation of the measures

18.31 We propose to define 'age assessment appeals' as a complaint by a UK user whose experience on the service is impacted because measures used to comply with illegal content safety duties (for example, the proposed livestreaming measures referred to above) have resulted in an incorrect assessment of the user's age.[604]

Appropriate action for age assessment appeals (services that are large or multi-risk)

18.32 We propose to recommend that services which are large or multi-risk should:

a) Have regard to the following matters in determining what priority to give to consideration of an age assessment appeal:

   i) The seriousness of the impact on the user as a result of the assessment of their age;
   ii) Whether the decision made on the basis of the user's age was made without human oversight and, if so, information available about the accuracy of the specific technology used in making age assessments of the type concerned;
   iii) The past error rate on the service in relation to age assessments of the type concerned; and
   iv) Any representations made by the user as part of the complaint as to the effect of the decision on their livelihood;

b) Monitor their performance against performance targets relating to:

   i) The time it takes to determine the age assessment appeal; and
   ii) The accuracy of decision making;

   c) and should resource itself to give effect to those targets;

d) If the provider determines that the user's age was incorrectly assessed, the provider should take any necessary steps to restore the user to the position they would have been in had the assessment been correct, so far as is appropriate and possible. For example, where an incorrect assessment of age led to a user not being able to receive comments while livestreaming, the provider should reverse this restriction so that in future the user is able to receive comments while livestreaming; and

e) The provider should monitor trends in age assessment appeals to help improve any age assurance process used on the service.

[604] As a consequence of introducing the defined term 'age assessment appeal', we are also proposing to change the existing defined term 'appeal' to 'content appeal'. This is necessary to distinguish it from an 'age assessment appeal'.

Appropriate action for age assessment appeals (services that are neither large nor multi-risk)

18.33 We propose to recommend that services which are neither large nor multi-risk should:

a) determine age assessment appeals promptly;
b) if the provider determines that the user's age was incorrectly assessed, the provider should take any necessary steps to restore the user to the position they would have been in had the assessment been correct, so far as is appropriate and possible. For example, where an incorrect assessment of age led to a user not being able to receive comments while livestreaming, the provider should reverse this restriction so that in future the user is able to receive comments while livestreaming; and
c) the provider should monitor trends in age assessment appeals to help improve any age assurance process used on the service.

Differences with equivalent measures in the Protection of Children User-to-user Codes

18.34 These measures are similar to PCU D11 and D12 in the Protection of Children User-to-user Codes. The key difference is that in PCU D11 and D12 the definition of 'age assessment appeals' and the description of the action that the provider must take if the appeal is upheld refer to restricting access to content.[605] However, the measures proposed in this consultation that involve the use of highly effective age assurance are broader, as they involve:

a) Restriction of functionalities (e.g. under the livestreaming measures a person assessed to be a child could not receive comments).
b) Settings being off by default (e.g. under ICU F1.6, where a service displays automated location information, it should implement default settings on the account of a person assessed to be a child ensuring that the location information associated with that account is not visible to other users of the service).
c) The provision of supportive information in certain circumstances, including when a person assessed to be a child seeks to change certain default settings.

18.35 For these reasons, it would not be appropriate to refer to restricting access to content in the definition of 'age assessment appeals' and in the description of the action that the provider must take if the appeal is upheld. Instead, we have:

a) Defined 'age assessment appeals' by reference to a user whose experience on the service is impacted because measures used to comply with illegal content safety duties have resulted in an incorrect assessment of the user's age'.
b) Described the action that the provider must take if the appeal is upheld as taking any necessary steps to restore the user to the position they would have been in had the assessment been correct, so far as appropriate and possible.

[605] The definition of 'age assessment appeals' is 'A complaint by a United Kingdom user who is unable to access content in or via the search results of the service because measures used to comply with a duty set out in section 29(2) or (3) of the Act have resulted in an incorrect assessment of the user's age'. PCU D12.3 provides that 'If the provider determines that the user's age was incorrectly assessed, the provider should take any necessary steps to enable the user to access content to which access was restricted as a result of that incorrect assessment (so far as appropriate and possible for the purpose of restoring the position to what it would have been had the assessment been correct)'.

Effectiveness

18.36 These measures ensure that users can appeal when they have been subjected to an incorrect age assessment (e.g. had functionalities restricted). An appeals mechanism helps to correct this. It does this by restoring the user to the position they would have been in if the provider assessed their age correctly, so far as possible and appropriate. This is also consistent with data protection laws to enable users a mechanism to request rectification of inaccurate personal data where an incorrect assessment of their age has occurred.

18.37 Recommending that providers monitor trends in complaints about incorrect assessments of a user's age will help to ensure its age assurance process is working effectively. For example, if a service provider using highly effective age assurance notices a surge in complaints about inaccurate age checks, it could investigate whether there is a problem with the accuracy of its chosen age assurance methods and take action to address this.

18.38 To the extent that we are proposing additional recommendations for large or multi-risk services, these relate to their prioritisation of age assessment appeals. Providers of services that are large or multi-risk may receive a higher volume of age assessment appeals and would therefore benefit from a prioritisation process and performance targets to address these appeals. The prioritisation process in this measure gives providers flexibility to set up their own process without causing operational difficulties, while also specifying the factors they should consider when establishing the process. This will aid their effective handling of these appeals.

Impacts and Costs

18.39 We have considered impacts on service providers who use highly effective age assurance for the purpose of targeting measures recommended in the Illegal Content User-to-user Codes. We note that all providers within the scope of this measure will also be within the scope of equivalent measures concerning appeals against action taken by providers against suspected illegal content.[606]

18.40 As the activities and costs involved in the measures are likely to be similar, we expect there to be some overlap in actions required and the associated costs.

18.41 Despite this, there is likely to be some increase in costs to implement this measure. This could be due to the increased number of appeals, additional costs associated with appeals related to age assurance (e.g. staff with expertise on age assurance), and monitoring the performance of age assurance decisions.

18.42 We would generally expect the volume of appeals a service provider receives to increase with the size of the service because larger services are likely to make a greater number of age assessment decisions. This means that the highest costs will likely be incurred by the providers of the largest services who are most likely to be able to absorb them, and who we expect would cause the greatest benefits from implementing these recommendations.

18.43 We have proposed additional prioritisation measures for large or multi-risk services. As noted above, we think these will aid their effective handling of age assessment appeals. These services are also already subject to a similar measure related to appeals (see ICU D8).

[606] Illegal harms measures ICU D8, ICU D9, ICU 10.

We consider the incremental cost associated with the equivalent age assessment appeal prioritisation process is likely to be limited.

18.44 We expect decisions on age assurance appeals to be more straightforward (and therefore less costly) than other appeals on whether content is illegal (or an illegal content proxy), which are likely to be more diverse and subjective. In addition, we are also providing flexibility to services in the length of timescales they need to adhere to in resolving complaints. Finally, we note that service providers may already have mechanisms for users to complain about inaccurate personal data for the purpose of complying with data protection law.

18.45 Overall, we consider the additional costs identified above are likely to be small, relative to the benefits the measure will provide in increasing user satisfaction, ensuring the age assurance process is highly effective, and that user rights to freedom of expression and privacy are protected.

Rights assessment

Freedom of expression and freedom of association

18.46 As noted above, these proposed measures are similar to PCU D11 and PCU D12 in the Protection of Children User-to-User Code. We assessed the rights impacts of those measures in our April 2025 Statement.[607] There, we concluded that that any interference with freedom of expression is likely to be limited. Our view is the same for these new proposals: we do not consider that providers taking appropriate action in response to age assessment appeals will interfere with individuals’ rights to freedom of expression and association.

18.47 Instead, we consider they could have positive impacts on the right to freedom of expression. This is because the measures are aimed at service providers remedying restrictions which have been applied to user’s accounts based on an incorrect assessment of age. For this reason, we refer to these measures as safeguards in the Illegal Content User-to-user Codes in relation to our proposals on livestreaming and grooming.

Privacy and data protection

18.48 These measures could have a positive impact on the right to privacy by providing greater transparency and accountability around decisions that are made in relation to incorrect assessments of age. We consider that the proposal will incentivise providers to reduce the number of incorrect decisions taken due to the resources and time required to consider complaints.

18.49 Providers should ensure they comply with data protection laws and familiarise themselves with any relevant guidance issued by the ICO. We have recommended specific measures to enable users to appeal against an incorrect assessment of age where highly effective age assurance is used. This ensures a clear mechanism through which individuals can safeguard the accuracy of their personal data.

18.50 We therefore consider that the impact of the measures on individuals’ rights to privacy (and data protection) to be relatively limited, and to have the potential to be positive overall.

[607] Ofcom, 2025, Statement: Protecting Children from Harms Online from paragraph 16.232.

What providers should implement the measure

18.51 We consider that the measures should apply to services that use highly effective age assurance to determine which UK users of the service are children for the purpose of targeting measures recommended in the Illegal Content User-to-user Codes at such users, their user accounts or their content feeds. This is because the risk that a user may be impacted by an incorrect assessment of age will only arise in this circumstance.

Proposal to make minor amendments to the Part 3 HEAA Guidance

18.52 In May 2024, as part of our consultation on protecting children online, we consulted on non-statutory draft guidance for highly effective age assurance to assist service providers in scope of Part 3 of the Act in complying with their duties. Subsequently, in January 2025 we published the final Part 3 HEAA Guidance. As of May 2025, we have updated the Part 3 HEAA Guidance with references to the Protection of Children User-to-user Codes and to reflect any changes to the wording and numbering of the relevant measures.

18.53 Service providers in scope of Part 3 of the Act should refer to this guidance to understand what constitutes highly effective age assurance and how it should be implemented for the measures set out in the Illegal Content User-to-user Code.

18.54 We are now proposing highly effective age assurance for the purpose of certain measures in the Illegal Content User-to-user Codes. Therefore, we propose amendments to the Part 3 HEAA Guidance to reflect this expansion of scope by adding references to the Illegal content User-to-user Codes and to reflect any changes to the wording and numbering of the relevant measures. This will enable the Part 3 HEAA Guidance to assist Part 3 services in complying with both sets of codes and future-proofs the guidance for use as our Codes evolve.

18.55 We set out a markup at Annex 10.

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

Summary

In our Illegal Content Codes of Practice we recommend a set of default safety settings for child user accounts (ICU F1) and supportive messaging features (ICU F2) for child users to make it harder for perpetrators to identify and interact with child users. This includes settings that make it harder for adults to contact children they don't know.

We had always intended that service providers should use highly effective age assurance to implement these measures. However, when we published our first Illegal Content Codes of Practice, we were not in a position to include this recommendation because we had not published our Guidance on Highly Effective Age Assurance for Part 3 Services (Part 3 HEAA Guidance). Now that we have done so, we are proposing that providers either:

Option A: Implement highly effective age assurance to apply the safety settings and supportive messaging to all users that have not been determined to be an adult.

Option B: Apply the safety settings and supportive messaging to all users of the service.

Giving providers two options to apply these measures allows them to select the one which works best for their service, while ensuring a higher level of protection for children.

Our Proposals

Number	Proposed measure	Who should implement this
ICU F1 ICU F2	We are recommending amendments to ICU F1 and F2. These proposed changes would mean that services in scope of ICU F1 and F2 should implement these measures using one of the following approaches: • Using highly effective age assurance, as set out in the Part 3 HEAA Guidance; or • Applying ICU F1 and F2 to all users of the service	This measure is an amendment of the measures ICU F1 and F2, therefore our proposals will apply to the same set of services as set out in our December 2024 statement: • All regulated user-to-user services at high risk of grooming • Regulated services that have at least a medium risk of grooming, and are a large service (7 million or more monthly UK users)

Consultation questions

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

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

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

Introduction

19.1 Online grooming involves a perpetrator communicating with a child with the intention of sexually abusing them, either online or in person. Children may also experience other forms of sexual abuse offences online as part of the grooming process, including sexual communication with a child and causing or inciting a child to engage in a sexual act. The scale of online grooming is significant, and perpetrators will deploy a range of tactics to manipulate, coerce or incite a child into sexual activity.[608]

19.2 Grooming will affect each victim and survivor differently, but the effects are significant and can last a lifetime. Individuals who have experienced online grooming report negative psychological impacts such as self-harming, loss of confidence, aggression, self-blame and lack of personal trust. The Illegal Harms Register of Risks (the Register) provides a full overview of what grooming is, the impacts and how it manifests online.

19.3 To address this harm, measures ICU F1 and F2 in the Illegal Content User-to-user Codes of Practice (Illegal Content U2U Codes) are designed to combat grooming for the purpose of child sexual exploitation and abuse (CSEA) on user-to-user services. They are:

- Measure ICU F1: Safety defaults for child users - This includes measures to remove child users from network expansion prompts, ensure child users are not included in connection lists and ensure users are not able to direct message child users without an established connection.

- Measure ICU F2: Support for child users - This includes supportive information at four key user journey points, including when children are turning off defaults, establishing a new connection with a user, sending a direct message to user for the first time, and taking action against an account.

19.4 ICU F1 are safety settings that will restrict the use of specific functionalities; children will be able to turn them off. They aim to prevent adults from finding and communicating with children, which are crucial steps in the commission of grooming offences. ICU F2 recommends that service providers give child users supportive information at key points of the user journey that should be prominently displayed and comprehensible. We are not reconsulting on our position as set out for these measures in our December 2024 Illegal Harms Statement (December 2024 Statement).

19.5 Our December 2024 Statement predated the publication of Part 3 HEAA Guidance. At the time of publication, we could therefore not link the safety default measures to highly effective age assurance. Rather than delay the introduction of the measures until we had finalised our policy on highly effective age assurance, we set out in the December 2024 Statement that as an interim measure a provider should apply ICU F1 and F2 'to the extent that it has an existing means to determine the age or age range of a particular user of the service' and that it should use its existing means of determining age to do so.

[608] Ofcom, 2024. The Causes and Impacts of Illegal Harms (Register of Risks). [accessed 11 June 2025]

19.6 We recognised that this was suboptimal as most services rely on self-declaration of age, which is not capable of being highly effective at correctly determining whether or not a user is an adult. We therefore said that we would later consult on amending ICU F1 and F2 to extend the use of highly effective age assurance to these measures. It has therefore always been our intention that service providers should use highly effective age assurance for the purposes of applying the measures. Now we have finalised our approach to highly effective age assurance, we are doing so.

19.7 As set out below, we are also proposing that, as an additional option to using highly effective age assurance, service providers should have the option of applying the measures to all their users. As we explain below, we consider that this would be equally effective at protecting children.

Our proposals

19.8 We are recommending amendments to ICU F1 and F2. These proposed changes would mean that services in scope of ICU F1 and F2 should implement these measures using one of the following approaches:

- Using highly effective age assurance, as set out in the Part 3 HEAA Guidance; or
- Applying ICU F1 and F2 to all users of the service[609].

19.9 In this chapter, we set out our proposals for how these approaches to implement ICU F1 and F2 will ensure that children are more effectively protected from grooming.

19.10 We are also proposing to remove the existing provision that in scope services should apply ICU F1 and F2 'to the extent that it has an existing means to determine the age or age range of a particular user of the service'. We assess that this change will ensure more services are implementing ICU F1 and F2, as services will need to either implement highly effective age assurance or apply the measures to all users, and therefore increase the number of services upon which children benefit from the protections of the measures.

Interaction with the regulatory framework

19.11 CSEA is a priority harm under the Online Safety Act 2023 (the Act). We have set out measures to mitigate the risk of CSEA on user-to-user services, specifically child sexual abuse material (CSAM) and online grooming, in our December 2024 Statement.

19.12 Since December 2024, we have published:

- the January 2025 Age Assurance and Children's Access Statement;
- the Part 3 HEAA Guidance; and
- the April 2025 Statement on Protecting Children from Harms Online (April 2025 Statement)[610].

[609] Note that users need to be 'logged in' to their account to access the functionalities affected by the measures. Therefore, any logged out users would not see any changes to their use of the service.

[610] As part of the April 2025 publication, we updated the Part 3 HEAA Guidance with references to the Protection of Children U2U Codes and to reflect any changes to the wording and numbering of the relevant measures.

19.13 There are some services that are in scope of both ICU F1 and F2 as well as the measures in the Protection of Children U2U Codes that require the use of highly effective age assurance. We expect that service providers in scope of the children's safety duties that require them to implement the age assurance measures as set out in the Protection of Children U2U Codes will use this information to target ICU F1 and F2.

19.14 Services in scope of ICU F1 and F2 that are not in scope of other protection of children measures that require the implementation of highly effective age assurance, could choose to implement highly effective age assurance to target ICU F1 and F2, or apply them to all users.

19.15 Where service providers are recommended to implement highly effective age assurance to target ICU F1 and F2, we expect providers to have regard to certain principles when implementing an age assurance process[611]. We explain this measure in detail in Chapter 18.

Option A: Applying Highly Effective Age Assurance to target ICU F1 and F2 at children

19.16 Historically, many service providers have used self-declaration of age to determine which of their users are adults. However, this is not a reliable way to determine which users are children or adults. Our research shows that as many as one in five 8-17 year olds present with false ages to appear 18 and over[612]. This means that without the changes we are proposing in this consultation a substantial proportion of children would not benefit from the protection of the measures.

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

19.18 We consider our approach to highly effective age assurance to be flexible, tech-neutral and future-proof, with the protection of children at its heart. We set out the criteria for an age assurance process to be considered highly effective in our Part 3 HEAA Guidance and explain our approach in more detail in Chapter 18.

Option B: Applying ICU F1 and F2 to all accounts

19.19 We are proposing that where a service provider does not implement highly effective age assurance, they should apply ICU F1 and F2 to all users. Applying ICU F1 and F2 to both children and adult users would mean that all children – as well as all adults – are provided with the protections offered by ICU F1 and F2, thus increasing the protections against online grooming.

19.20 Similar to option A, this option also ensures that services are not relying on self-declaration as a means to apply the measures to children. In this case, regardless of how a provider will determine a user's age, all users will benefit from the protections of ICU F1 and F2. In turn, this will ensure that materially more children than currently expected, benefit from the protections afforded by the measures. This would contribute to the reduction of online grooming, thereby delivering significant benefits.

[611] See ICU B1 – Annex 7 for further details.

[612] Ofcom, 2022. Children's Online User Ages Quantitative Research Study. [accessed 11 June 2025]

19.21 This option could also make it harder for perpetrators to identify and interact with children they do not know. For example, all users – both children and adults – will not be presented with network expansion prompts as a default, which could make it more difficult for perpetrators to connect with new accounts and therefore make it harder for them to identify children's accounts online.[613]

Impacts and Costs

19.22 We expect that some service providers in scope of ICU F1 and F2 in the Illegal Content U2U Codes may also be in scope of children's safety duties and relevant measures in the Protection of Children U2U Codes, which require the use of highly effective age assurance. In these cases, services can apply existing highly effective age assurance processes for the purpose of implementing ICU F1 and F2. We expect there would be relatively small additional costs for these services to implement this measure, relative to the initial costs of implementing highly effective age assurance on a service.[614]

19.23 There may be some service providers that (1) are not in scope of the children's safety duties and relevant Protection of Children U2U Codes[615] and (2) do not already use highly effective age assurance. For example, certain gaming, messaging or social media services may be high or medium risk for grooming but may not use highly effective age assurance.

19.24 In this section, we assess costs for providers that do not already use highly effective age assurance. Under our proposals, they should either implement:

- Option A: Apply highly effective age assurance; or
- Option B: apply ICU F1 and F2 to all users

19.25 The direct costs may significantly vary depending on the size and nature of the service and how providers choose to implement measures ICU F1 and F2.

Option A: Costs of applying highly effective age assurance

19.26 We analysed the costs of highly effective age assurance in the April 2025 Statement. We have used the same analysis here, using the same data, which we summarise in this section.[616]

19.27 Providers who opt to apply highly effective age assurance will likely incur a variety of costs in the process as discussed in the following paragraphs.

Direct costs

19.28 Providers are likely to incur some one-off preparatory costs. These may include costs to familiarise staff with the measures and guidance, and to understand, assess and choose among various options available to the provider.

19.29 We estimate direct costs would be incurred either through use of a third-party provider of age assurance services or through development of an in-house age assurance capability.

[613] Ofcom, 2024. The Causes and Impacts of Illegal Harms (Register of Risks). [accessed 11 June 2025]

[614] We would expect some small costs associated with linking any existing age assurance process to the functionalities associated with ICU F1 and F2.

[615] For example, because they prohibit all kinds of primary priority content (PPC) and priority content (PC).

[616] See Annex 15 for further details.

Third-party direct costs comprise the per age check costs from the third-party provider of age assurance services and would depend upon how many users the service needs to check.

Cost of third-party provision

19.30 Where a provider chooses to use third-party age assurance methods, the main cost component would be the per-check cost. These per-check costs are likely to vary depending on the age assurance process and provider, as underlying costs and pricing approaches vary.

19.31 In addition, there may be upfront costs linked to (1) the age assurance provider setting up a client account to prepare the age assurance method for use; (2) the service provider having to introduce changes to the existing ICT infrastructure or building of a new user interface to integrate age assurance with the service; and (3) the service having to train some of its staff when the process becomes operational.

19.32 To illustrate what these costs may mean for a service, we set out cost examples for hypothetical services with different numbers of users in Table 19.1. We assume that:

- There is a fixed cost per check of between £0.05 and £0.30;
- Each user in the entire existing user base is checked once leading to a one-off cost;
- The cost per check will remain unchanged over time; and
- Each new user will be checked once and these ongoing age checks on new users will continue annually.

Table 19.1: Illustrative cost estimates of age checks via third-party age assurance providers*

Existing UK user base	New users each year	Age assurance for existing users (one-off)	Age assurance for new users (annual ongoing cost)
1,000	100	£50 - £300	£5 - £30
10,000	1,000	£500 - £3,000	£50 - £300

Existing UK user base	New users each year	Age assurance for existing users (one-off)	Age assurance for new users (annual ongoing cost)
100,000	10,000	£5,000 - £30,000	£1,000 - £3,000
350,000	35,000	£18,000 - £105,000	£2,000 - £11,000
700,000	35,000	£35,000 - £210,000	£2,000 - £11,000
1,000,000	50,000	£50,000 - £300,000	£3,000 - £15,000
7,000,000	70,000	£350,000 - £2,100,000	£4,000 - £21,000
20,000,000	200,000	£1,000,000 - £6,000,000	£10,000 - £60,000

Source: Ofcom analysis

**Note: For existing UK user base of 100,000 and more, cost estimates have been rounded up to the nearest thousand, while for existing UK user base of 1,000 and 10,000 cost estimates turn out to be small and have been presented without any rounding. These illustrative 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).*

19.33 These estimates suggest smaller services with up to 700,000 initial users requiring age checks may incur one-off costs of up to £210,000 for existing users. They may also incur annual ongoing costs of up to £11,000 to check their new users.

19.34 Larger services with up to 20 million initial users requiring age checks may incur upper range one-off costs of up to £6 million for existing users and up to £60,000 in annual ongoing costs to check new users. However, larger services may be able to negotiate lower per-check costs based on volume, which means this upper range may be overestimated.

19.35 In practice, some services may implement highly effective age assurance by making it a voluntary check for users to determine whether they are an adult and thus conduct age checks only on a subset of all users. Costs could be lower in such cases.

Costs of in-house highly effective age assurance provision

19.36 If a service chooses to build its own age assurance method, this would require significant upfront investment required to build an in-house platform which enables age assurance for users.

19.37 Developing an in-house age assurance method would entail staff costs associated with developing the age assurance software platform and the ongoing staff costs involved in its operation. These costs are estimated to be up to £1 million in platform development costs and £1 million annually in ongoing staff costs.

19.38 Based on the substantial estimated upfront costs relating to developing an age assurance software platform, we expect that any services seeking to develop age assurance methods in-house are likely to have many users requiring age assurance. This option may be more cost-effective if a service (1) predicts a high number of new service users over time and (2) expects the ongoing costs to be lower than those of ongoing age checks by a third party.

Indirect costs

19.39 Some users of services with highly effective age assurance may not want to go through the age assurance process due to concerns around privacy and data protection.[617] This may lead to services losing traffic if they make highly effective age assurance obligatory for all users, particularly if users see limited benefit from providing their data to the service. This is likely to be less of an issue if highly effective age assurance is a voluntary option, as users will have an alternative of having safety measures ICU F1 and ICU F2 applied to their account. There are likely to be some indirect costs to adult users from these measures as they are likely to increase friction and impact on service functionality (as described in paragraph 19.31). The overall magnitude of any indirect impact on users would therefore depend on how much either age assurance or the application of ICU F1 and F2 diminishes the user's experience of the service.

19.40 In practice, it is challenging to quantify these types of effects. There is no evidence to suggest which of these impacts would be more substantial. At this stage it is unclear what users may choose to do if highly effective age assurance is implemented by a service.

Option B: Cost to apply ICU F1 and F2 to all users

Direct costs

19.41 If services choose to apply the measures to all users, they could avoid incurring the costs of implementing highly effective age assurance for ICU F1 and F2.

19.42 Some direct costs may be incurred by providers when implementing ICU F1 and F2 for all users. In the December 2024 Statement, we estimated that the total one-off upfront cost could range from £10,000 to £325,000 while annual ongoing costs could range from £2,500 to £81,250.[618] We expect that any additional costs on top of those set out in our December 2024 Statement would be small or negligible.[619]

Indirect costs

19.43 In the December 2024 Statement,[620] we set out that services would incur indirect costs from applying ICU F1 and F2 to children's accounts, as it results in 'lower user activity and subsequent lost revenue'. We assessed that there could be a reduction in legitimate activity on a service, which could have an indirect cost, for example, if users spend less time on a service leading to potential loss in advertising revenue for the provider.

19.44 If a service decides to apply the measures to all users rather than using highly effective age assurance, indirect costs on the service would increase because ICU F1 and F2 would affect adult users. We discuss these in the following paragraphs.

[617] Ofcom, 2025. Protecting Children from Harms Online (Volume 5), p.12, paragraph A3.47. [accessed 12 June 2025]

[618] Ofcom, 2024. Service Design and User Choice (Volume 2), pp.369-370, paragraph 8.89-8.95. [accessed 11 June 2025]

[619] Some additional costs may be required to ensure the measures are applied to all users, not just child users.

[620] Ofcom, 2024. Service Design and User Choice (Volume 2), p.370, paragraph 8.96-8.97. [accessed 11 June 2025]

Impacts on users

Adverse effects on users

19.45 This proposed measure will mean that ICU F1 and F2 will now be applied to all users, both children and adults. This means that the restrictions will be applied to more accounts.

19.46 We previously considered the adverse impacts on children subject to ICU F1 and F2 in the December 2024 statement.[621] There are likely to be some adverse impacts on child users due to a more limited functionality, however we consider this would be outweighed by the benefits from a lower risk of grooming to those children. As set out above our analysis suggests that the benefits for children outweigh the reduced functionality.

19.47 For adult users, the adverse effects would be related to the inconvenience of highly effective age assurance or the application of the measures to their account.

19.48 For service providers that adopt Option A: (apply highly effective age assurance), if a service provider makes it a requirement for all users, then the adverse effect on adults from highly effective age assurance would be whatever inconvenience is caused by having to go through this process (including having to provide some personal data). If a service provider applies highly effective age assurance on a voluntary basis, then adult users will have to choose between undergoing age verification or having default safety settings applied to their account (ICU F1) and supportive information that cannot be turned off (ICU F2).

19.49 For service providers that adopt Option B: (apply ICU F1 and F2 to all users), adult users would be adversely affected by the inconvenience of having default safety settings applied to their account (ICU F1) and supportive information that cannot be turned off (ICU F2).

19.50 Users wanting to change the default settings will be able to do so manually but will not be able to turn off the supportive information prompts, unless they go through the highly effective age assurance process, if available. Also, while any individual user will be able to turn off the default safety settings (ICU F1) on their own account, we do not expect all adult users to do so. This may have some negative impact on potential network benefits to adult users due to a reduction in the pool of potential connections to other adult users. Some adult users may also experience benefits from the measures, such as increased privacy.

19.51 Finally, some services may make completing the highly effective age assurance process a requirement for all users (for example, by making users undergo highly effective age assurance at the point of sign up or log in). This would also have an impact on children because they must establish their age to access the service (or certain features on the service, depending on the approach taken by the provider). This would impact the child's user experience and will require that the child provides some personal data, the exact nature of which depends on the type of highly effective age assurance solution used.

Rights assessment

19.52 We acknowledge that the introduction of an age assurance recommendation to ICU F1 and F2 may impact a users' rights to freedom of expression and association and has potential implications for their privacy. The degree of interference with these rights will be subject to

[621] Ofcom, 2024. Service Design and User Choice (Volume 2), pp.370-371, paragraph 8.98-8.104; p.392, paragraph 8.192-196. [accessed 11 June 2025]

a provider's decision as to whether they implement highly effective age assurance (and the type of highly effective age assurance they decide to implement) or apply the measures to all users. We therefore consider each of these scenarios separately below.

19.53 We note that there could be some interference with adult users' rights where our recommendation may make them subject to ICU F1 and F2 (for example, where adult users do not confirm their age using highly effective age assurance and thus the service assumes they are a child user, or on services where the measures are applied to all users). In the December 2024 Statement, we assessed the impact on adults of ICU F1 and F2 being applied to child user accounts, noting that this could result adults being restricted from legitimate engagement with child users. However, we have not assessed the impact on adults in cases where ICU F1 and F2 are applied to both adults and child user accounts, as is proposed here. We discuss this potential further level of impact due to our proposal below.

Freedom of expression and association

On services where highly effective age assurance is implemented

19.54 ICU F1 and F2 will apply to adult users in cases where an adult chooses not to complete an age check (such as where highly effective age assurance is implemented in a way that makes it optional for users). In the December 2024 Statement[622], we set out that the impact of ICU F1 and F2 on users would be that they were unable to form connections and engage through the relevant functionalities. However, we also explained that this impact is negligible because the ICU F1 settings can be manually switched off and because ICU F2 has a negligible impact on a user's rights. We consider that our proposals will have a similar negligible impact on adult users' freedom of expression where ICU F1 and F2 are applied.

19.55 Our proposal will broaden the scope of ICU F1 and F2 by removing the application requirement of 'existing means'. This will impact children by making more children subject to ICU F1 and F2. However, we consider this to be proportionate (as set out in the December 2024 Statement). The impact on child users' freedom of expression is mitigated through the default nature of the settings and, in any case, is proportionate to address the risk of grooming on relevant services. Despite the limited impact on users' rights from the introduction of highly effective age assurance, we acknowledge that a new requirement by services on users to complete an age check may have some other impacts on users' rights.

19.56 The age assurance process may create friction in the user experience and may result in some adults choosing not to use a service, which may consequently restrict their ability to use a service's functionalities to impart and receive information. This impact is limited because we consider it is a user's choice to complete an age check and because service providers will be incentivised to ensure that the method of age assurance they choose minimises friction to the user experience. Our recommendation for providers to have regard for principles of accessibility and interoperability when implementing an age assurance process will also limit the impact of our proposal on user experience.

19.57 We note that there is a risk that the age assurance process may incorrectly assess some adults to be a child and subject them to ICU F1 and F2. However, this will likely create minimal friction as these adult users will still be able to turn off the settings for ICU F1 and because the impact of ICU F2 is negligible. Furthermore, we consider that while there is

[622] Ofcom, 2024. Service Design and User Choice (Volume 2). [accessed 11 June 2025]

potential risk for a margin of error in the use of highly effective age assurance, this risk will be limited if providers take account of the recommendations in our Part 3 HEAA Guidance. This sets out criteria and recommendations to ensure the age assurance process is highly effective. Such recommendations include adopting processes which are technically accurate, robust, reliable and fair, with the aim of ensuring that incorrect assessments are avoided and, where applicable, remedied efficiently.

On services that apply the ICU F1 and F2 measures to all users

19.58 Where services apply the measures to all users, adult users will have their user experience changed. All users (both adults and children) will be subject to the safety defaults and support for child users' measures (ICU F1 and F2), which may have an impact their freedom of expression and association. However, we consider this impact to be minimal because adult users will have the option to turn off ICU F1 settings and because ICU F2 has a negligible impact on freedom of expression rights.

Privacy and data protection

19.59 Article 8 of the European Convention on Human Rights (ECHR) sets out the right to respect an individual's private and family life. The use of an age assurance process to determine if a user is an adult will involve the collection and processing of personal data.

On services that implement highly effective age assurance

19.60 All methods of age assurance involve the processing of some personal data of individuals. Therefore, where highly effective age assurance is implemented, it will have an impact on users' rights to privacy and their rights under data protection law. However, we consider that users' privacy rights will not be disproportionately impacted because service providers are required to comply with relevant data protection legislation when processing any personal data.

19.61 The degree of interference will depend on a number of variables, including the nature of the information required to complete the age assurance process. For example, if more sensitive information is required, the interference may be greater. We are clear in our Part 3 HEAA Guidance[623] and our January 2025 Statement[624] that in implementing a highly effective age assurance process, services are bound by data protection laws. Compliance by service providers with both the online safety and the data protection regime is mandatory and should not be considered a trade-off. As we state in the Part 3 HEAA Guidance[625], service providers should consult the Information Commissioner's Office (ICO) guidance to understand how to comply with the data protection regime.

On services that apply ICU F1 and F2 to all users

19.62 Where a provider chooses not to implement highly effective age assurance and instead apply ICU F1 and F2 to all users, we consider that there will be no significant impact on users' rights to privacy or their rights under data protection. We do not expect providers to

[623] Ofcom, 2025. Guidance on Highly Effective Age Assurance for Part 3 Services. [accessed 16 June 2025]

[624] Ofcom, 2025. Statement: Age Assurance and Children's Access. p.10, paragraph 2.23. [accessed 16 June 2025]

[625] Ofcom, 2025. Guidance on Highly Effective Age Assurance for Part 3 Services. p.22, paragraph 5.5. [accessed 16 June 2025]

collect any further personal data or information about a user to implement the measure via this approach.

Which providers should implement this measure

19.63 This proposal is an amendment to ICU F1 and F2 to increase the effectiveness of the measures set out in our December 2024 statement. As a result of our amendments, we expect our proposals to apply to a wider set of services[626] and relevant service providers will be asked to identify their risk level using the risk assessment, outlined in our December 2024 statement.

19.64 Once a service provider has completed the risk assessment based on our December 2024 Statement, providers will be recommended to implement either Option A or Option B if they identify as one of the following:

- Regulated user-to-user services at high risk of grooming; or
- Regulated user-to-user services that have at least a medium risk of grooming and are a large service (7 million or more monthly UK users).

19.65 In both cases, the measure will only apply to regulated user-to-user services that have the relevant functionalities targeted by the measures. These are network expansion prompts (ICU F1.2), connection lists (ICU F1.3), direct messaging (ICU F1.4), and automated location information display (ICU F1.6).

Provisional conclusion

19.66 As we have explained in this chapter, recommending that providers in scope of ICU F1 and F2 either use highly effective age assurance to target the measures at users that have not been determined to be adults or apply the measures to all users, would materially increase the number of children that benefit from the protections from the measures set out in our Illegal Content U2U Codes. Therefore, we consider our recommendations could result in material reductions in incidents of grooming. Given the prevalence and severity of grooming, we consider that the associated benefits would be very significant.

19.67 Large services could incur very significant costs should they choose to comply with our proposal by applying highly effective age assurance. However, given the severity of the harm we are attempting to tackle with the proposal, we provisionally consider that these costs would be proportionate. Our view that the proposal in this chapter is proportionate is reinforced by the facts that:

- The costs could be scaled to the size of individual services due the existence of third-party providers providing highly effective age assurance on a per check basis.
- Services with fewer than 7 million users will only be in scope of the proposal where they pose a high risk of grooming.
- Some service providers are likely to already be in scope of the Protection of Children U2U Codes that require the use of highly effective age assurance. We consider that the cost to these services to comply with our proposal would be relatively small.

[626] See paragraph 19.10 for details on amendments to expand the number of services ICU F1 and F2 are recommended to.

- Service providers have the option of complying with our proposal by applying the grooming protections to all users. As we explain in paragraph 19.42, the direct costs of doing this would be small.

19.68 Based on our assessment, we consider it proportionate to recommend that providers of services in scope of ICU F1 and F2 either:

- Implement highly effective age assurance to apply ICU F1 and F2 to all users that have not been determined to be an adult.; or
- Apply ICU F1 and F2 to all users of the service.

20. Crisis Response

Summary

During a crisis, certain kinds of illegal content and/or content harmful to children can spread rapidly online. In some cases, this can create significant risks to public safety in the UK.

Evidence from previous crises has shown how perpetrators use online services in a variety of ways to carry out illegal activity such as inciting racial or religious hatred, making threats, or inciting violence. Not only can this lead to an increase in the amount of illegal content circulating online but it can result in violence offline.

This was demonstrated by the nationwide violent and hateful disorder that followed the murder of three young girls in Southport in 2024. This led to – and was in turn facilitated by – the posting of illegal content . The resulting harm was severe and widespread, with attacks on people and property on the basis of ethnic and religious hatred.

Such crises are exceptional, and this means that online service providers' usual moderation measures may not be sufficient in these circumstances. We are therefore proposing a set of crisis response measures.

The first element of these measures recommends that providers have a crisis response protocol in place. The crisis response protocol should include monitoring indicators, to identify when a crisis has been initiated, and set out how the provider will stand up a crisis response team. Providers of large services should have a dedicated channel by which a relevant law enforcement authority can contact them in a crisis.

The second element recommends that providers who have identified a crisis (as per our definition) conduct a post-crisis analysis.

The proposed measures complement other proposals in this consultation, which will be relevant during a crisis. These include hash matching for terrorism content, and measures to exclude content potentially indicated to be priority illegal content from recommender feeds.

We propose that the crisis response measures should apply to:

- large[627] user-to-user services that are at medium risk, and;
- user-to-user services of any size that are at high risk of any one of the following harms:
  - within priority illegal harms – terrorism, threats, abuse and harassment (including hate) and foreign interference;
  - within priority content harmful to children – abuse and hate, and violent content.

[627] In our Illegal Content user-to-user Codes of Practice we propose to define a service as 'large' as a service which has more than 7 million monthly active United Kingdom users (see paragraphs 5.7 to 5.10 of the Illegal Content Codes of Practice).

Our proposals

Number	Proposed measure	Who should implement this
ICU C15 / PCU C11[628]	The provider should prepare and apply an internal **crisis response** protocol. It should also conduct and record a post-crisis analysis. Providers of large services should implement a dedicated communication channel by which law enforcement can contact them on crisis-related matters during a crisis.	• Providers of large user-to-user services that are medium risk of relevant harms • Providers of all user-to-user services that are high risk of relevant harms

Consultation questions

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

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

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

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

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

Introduction

20.1 A crisis can exacerbate the risks of both online and offline harms. During a crisis, there may be an increase in both the volume of illegal and/or content harmful to children, and the risks such content poses in catalysing offline harm, including the risk that services will be used to commit and/or facilitate a priority offence.[629]

20.2 This phenomenon was illustrated by the 2024 UK riots following the Southport attack, where online content was followed by unrest and violence across the UK.[630] Following the murder of three girls on 29 July 2024 at a children's summer holiday dance class[631] in Southport, Merseyside, online services were used to spread hatred, provoke violence

[628] The measure we propose to include in our Protection of Children User-to-user Code will be limited to services likely to be accessed by children.
[629] As set out in Schedules 5, 6 and 7 of the Online Safety Act 2023.
[630] Institute for Strategic Dialogue, 2024. From rumours to riots: How online misinformation fuelled violence in the aftermath of the Southport attack. [accessed 5 May 2025]
[631] This incident resulted in the deaths of three children. Nine people were injured in the attack, six of them critically.

targeting racial and religious groups, and encourage others to attack and set fire to mosques and asylum seeker accommodation.[632]

20.3 In autumn 2025, we published our letter to the Secretary of State for the Department for Science, Innovation and Technology (DSIT), setting out our conclusions about the role of online services in the aftermath of the attacks.[633] In that letter we stated that:

- some tech firms reported to us that they took action to limit the spread of illegal content, although there was evidence that it still spread widely and quickly following the tragic murders in Southport;
- numerous convictions followed the violence, with some individuals involved convicted of posting online death threats or threats to cause serious harm,[634] stirring up racial hatred,[635] or sending false information with intent to cause harm.[636]
- there was a clear connection between online activity and violent disorder seen on UK streets; and
- if our Illegal Content and Protection of Children Codes had been in force at the time, we are confident that they would have provided a firm basis for urgent engagement with services on the steps they were taking to protect UK users from harm.

20.4 Following on from the work set out in this letter, we are proposing to add a set of measures to our Codes designed to improve service providers' responses to crises. We consider that our proposed measures complement our existing Codes by putting specific measures in place to ensure providers act promptly to reduce the increase in and spread of the relevant harms during a crisis.

20.5 Since crises unfold in stages,[637] our proposed measures are developed, so that service providers implement safety measures at each stage; from pre-crisis to post-crisis.

What risks do crises pose?

20.6 Research has shown that illegal and/or content harmful to children stemming from crises poses an imminent risk to people both online and offline. For instance, offline trigger events are often followed by increases in types of online hate on services of all sizes.[638]

[632] Spring, M., 2024. Did social media fan the flames of riot in Southport? BBC, 31 July. [accessed 5 May 2025].
[633] Ofcom, 2024. Letter from Dame Melanie Dawes to the Secretary of State, 22 October 2024. [accessed 5 May 2025].
[634] Burnell, P., 2025. Taxi driver who stoked Southport riots jailed. 6 January. [accessed 5 May 2025].
[635] Comerford, R., 2024. Men jailed for encouraging unrest on social media. 9 August. [accessed 5 May 2025].
[636] The Guardian, 2024. Chester woman, 55, arrested over false posts about Southport murders. 9 August [accessed 5 May 2025].
[637] Porot, G., 2024. What are the Three Stages of Crisis Management?, International SOS, 19 February. [accessed 5 May 2025].
[638] Lupu, Y., Sear, R., Velásquez, N., Leahy, R., Restrepo, NJ., Goldberg, B., et al., 2023. Offline events and online hate. *PLoS ONE 18(1): e0278511*. [accessed 5 May 2025].

20.7 Past cases have shown that online terrorist content can increase exponentially following crises and related offline events. An increase in online terrorist content can also lead to offline violence.[639] [640]

20.8 When there is heightened uncertainty and fear – emotions that often arise during a crisis – this may manifest online. This may result in groups of individuals who share a similar identity (be it racial, ideological, or political) directing online hate, harassment, threats and abuse towards groups or individuals, who do not share their identity.[641] [642]

20.9 In other cases, it may manifest through violent content in which perpetrators call for violence[643] and/or depict imagery of violence.[644]

20.10 In rare cases, foreign actors may generate domestic unrest and violence by amplifying content connected to a crisis. Foreign interference operations can have longer term goals, such as using disinformation to foster an environment of distrust, casting doubt on the true accounts of events and fuelling political extremism and 'wedge issues'.[645] In the case of the Southport riots, the UK Government opened an investigation into the role of foreign interference in fostering the offline violence.[646] [647]

20.11 Having taken account of the evidence set out above and based on our assessment of past crises, we consider that the types of harms set out from paragraphs 20.7 to 20.10, and the content associated with them, are more likely to be relevant in a crisis than others. We think the types of harms, and associated content, that are most likely to increase and spread in a crisis, and therefore pose immediate risk to UK users, are the following:

- Certain kinds of priority illegal harms:
    - **Hate**:[648] content which may amount to hate offences listed in the Online Safety Act 2023 (the Act), such as content which displays abusive material that stirs up racial hatred or threatening written material intending to stir up religious hatred or hatred on the grounds of sexual orientation.
    - **Terrorism**:[649] content which may amount to terrorism offences listed in the Act, such as content which invites support for a proscribed organisation or to arrange a meeting supportive of a proscribed organization.

[639] Conway, M., Scrivens, R., and Macnair, L., 2019. Right-Wing Extremists' Persistent Online Presence: History and Contemporary Trends. *The International Centre for Counter-Terrorism – The Hague 10,* 1–24 [accessed 5 May 2025].

[640] For more information on the risks posed by Terrorism content, please see chapter 1 of our Illegal Harms Register of Risk. [accessed 5 May 2025].

[641] Council of Europe, 2023. Study on preventing and combating hate speech in times of crisis. [accessed 5 May 2025].

[642] For more information on the risks posed by hate, harassment, threats and abuse content, please see chapters 3 and 4 of our Illegal Harms Register of Risk. [accessed 5 May 2025].

[643] Gill, A., 2024. Mosque explosion call woman jailed for online post, BBC, 14 August. [accessed 5 May 2025]

[644] For more information on the risks posed by Violent content to children, please see chapter 7 of our Protection of Children Register of Risk. [accessed 5 May 2025].

[645] Intelligence and Security Committee of Parliament, 2020. Russia. [accessed 5 May 2025].

[646] Holden, M., Smout, Alistair., 2024. UK examines foreign states' role in sowing discord leading to riots, Reuters, August 5. [accessed 5 May 2025].

[647] For more information on the risks posed by Foreign Interference, please see chapter 16 of our Illegal Harms Register of Risk. [accessed 5 May 2025].

[648] See Chapter 3 of our Illegal Content Judgements Guidance [accessed 18 June 2025]

[649] See Chapter 2 of our Illegal Content Judgements Guidance [accessed 18 June 2025]

  - **Harassment, stalking threats and abuse**:[650] content which may amount to the harassment, threats, and abuse offences listed in the Act, such as content which makes a threat to kill or which involves behaving in a threatening or abusive manner likely to cause fear or alarm.
  - **Foreign interference offence**:[651] content which may amount to a foreign interference offence.
- Certain kinds of priority content harmful to children:
  - **Hate and abuse**:[652] content which incites hatred against people (1) of a particular race, religion, sex or sexual orientation, (2) who have a disability, or (3) who have the characteristics of gender reassignment; and content which is abusive and which targets any of the following characteristics: (1) race, (2) religion, (3) sex, (4) sexual orientation, (5) disability, or (6) gender reassignment.
  - **Violent content**:[653] content which encourages, promotes, or provides instructions for an act of serious violence against a person; and content which (1) depicts real or realistic serious violence against a person or (2) depicts the real or realistic serious injury of a person in graphic detail.

20.12 We have focused our proposed measures on these harms and refer to these harms as 'relevant harms' and the content associated with them as 'relevant illegal content' and/or 'relevant content harmful to children'.

20.13 To the extent that our proposed measures relate to illegal content, we are proposing to insert them into the Illegal Content User-to-User Codes.[654] To the extent they relate to content harmful to children, we are proposing to insert the measures into the Protection of Children User-to-User Code.[655]

Legal framework

20.14 The proposed measures are founded on service providers' existing duties under the Act. They are interconnected with and enhance existing content moderation measures as set out in the Illegal Content User-to-User Codes of Practice and the Protection of Children User-to-User Code of Practice (referred to collectively as Codes of Practice).

Illegal Content Safety Duties

20.15 Part 3 of the Act places duties on providers of regulated services. These include duties set out in section 10 of the Act that require providers of regulated user-to-user services to take or use proportionate measures relating to the design or operation of the service to (among other things): [656]

- prevent individuals from encountering priority illegal content;

[650] See Chapter 3 of our Illegal Content Judgements Guidance [accessed 18 June 2025]
[651] See Chapter 14 of our Illegal Content Judgements Guidance [accessed 18 June 2025]
[652] See Chapter 6 of our Guidance on Content Harmful to Children [accessed 18 June 2025]
[653] See Chapter 8 of our Guidance on Content Harmful to Children [accessed 18 June 2025]
[654] Ofcom, 2024. Illegal content Codes of Practice for user-to-user services [accessed 18 June 2025]
[655] Ofcom, 2025. Protection of Children Code of Practice for user-to-user services [accessed 18 June 2025]
[656] Section 10(2) to (3) of the Act.

- effectively mitigate and manage the risk of the service being used for the commission or facilitation of a priority offence;
- effectively mitigate and manage the risks of harm to individuals as identified in a service's most recent illegal content risk assessment; and
- minimise the length of time for which any priority illegal content is present and to swiftly take down illegal content when the provider becomes aware of it.

20.16 Priority illegal content is content that amounts to an offence specified in schedules 5, 6 or 7 of the Act (which includes threats, abuse and harassment (including hate), terrorism, and foreign interference).[657]

Children's Safety Duties

20.17 Part 3 of the Act also places duties on providers of regulated services likely to be accessed by children to take steps to prevent and protect children from encountering content harmful to children, including through content moderation where proportionate, and to take proportionate measures to effectively mitigate and manage the risk and impact of harm from content that is harmful to children.[658]

20.18 Priority content (PC) that is harmful to children includes, but is not limited to, hate and abusive content and violent content.

Our proposals

How the measures relate to our Codes of Practice

Interaction with the Illegal Content Codes and Protection of Children Code

20.19 It is important to note that our current Illegal Content User-to-User Codes and Protection of Children User-to-User Code include existing measures which address the relevant harms.

20.20 The Illegal Content User-to-User Codes and Protection of Children User-to-User Code recommend the following existing measures for user-to-user services, which will help to mitigate the risk of harm during a crisis:

- The Illegal Content User-to-User Codes recommend that services have systems and processes designed to swiftly take down illegal content and/or illegal content proxies of which they are aware, unless it is currently not technically feasible for them to achieve this outcome.[659] The Protection of Children User-to-User Code recommends that systems and processes are designed to swiftly action content which is harmful to children and/or harmful content proxy, that it is aware of, unless it is currently not technically feasible for them to achieve this outcome.[660] Thus, when a service becomes aware of such content during a crisis, it should be swiftly taken down, or take steps to prevent child users from encountering the content as applicable.
- The Illegal Content User-to-User Codes recommends that large or multi risk services should prepare a policy about the prioritisation of content for review, taking account of (among

[657] Section 59(10) of the Act.

[658] Section 12 (2) and (3) of the Act.

[659] See ICU C2 of our Illegal content Codes of Practice for user-to-user services [accessed 18 June 2025]

[660] See PCU C2 of our Protection of Children Code of Practice for user-to-user services [accessed 18 June 2025]

other things) the desirability of minimising the number of UK users encountering illegal content, and the severity of potential harm to UK users if they encounter illegal content.[661] This measure is replicated in the Protection of Children User-to-User Code in respect of content which is harmful to children.[662] Providers may decide to prioritise content for review during a crisis in a different way to at other times. By ensuring service providers are prioritising content for review in a rigorous way, these measures should help ensure that they deal with the most damaging pieces of content in an expeditious way during crises.

- The Illegal Content User-to-User Codes recommends that large or multi risk services should resource their content moderation function so as to give effect to internal content policies and performance targets, having regard to (among other things) the propensity for external events to lead to a significant increase in demand for content moderation.[663] This measure is replicated in the Protection of Children User-to-User Code in respect of content which is harmful to children.[664] These existing measures should help ensure that service providers are well placed to respond to surges in illegal content and content harmful to children during crises.
- The Illegal Content User-to-User Codes recommends that large or multi risk services should set performance targets for their content moderation function, and in doing so should balance the need to take relevant content moderation action swiftly against the importance of making accurate moderation decisions.[665] This measure is replicated in the Protection of Children User-to-User Code in respect of content which is harmful to children.[666] Once again, these existing measures should help ensure that service providers deal with illegal content and content harmful to children expeditiously in crises.

20.21 However, given the evidence we saw during the violent disorder following the Southport attack we consider that additional measures are necessary to address the risk of harm that may occur during a crisis faster and more effectively.

20.22 We are therefore proposing two new measures specifically relating to service providers' response to a crisis. These will work alongside and enhance existing measures.

Explanation of the measure

Definition of a 'crisis'

20.23 We propose to define 'crisis' for the purposes of these measures. We consider that this is important to ensure consistency across service providers in terms of what constitutes a crisis.

20.24 Our proposed definition of a 'crisis' in the Illegal Content User-to-User Codes is as follows:

[661] See ICU C5 of our Illegal content Codes of Practice for user-to-user services [accessed 18 June 2025]
[662] See PCU C5 of our Protection of Children Code of Practice for user-to-user services [accessed 18 June 2025]
[663] See ICU C6 of our Illegal content Codes of Practice for user-to-user services [accessed 18 June 2025]
[664] See PCU C5 of our Protection of Children Code of Practice for user-to-user services [accessed 18 June 2025]
[665] See ICU C4 of our Illegal content Codes of Practice for user-to-user services [accessed 18 June 2025]
[666] See PCU C4 of our Protection of Children Code of Practice for user-to-user services [accessed 18 June 2025]

- A crisis is an extraordinary situation in which there is a serious threat to public safety in the United Kingdom either:
  - as a result of a significant increase in **relevant illegal content** on the service; and/or
  - which has caused or is highly likely to cause a significant increase in **relevant illegal content** on the service.

20.25 We have defined '**relevant illegal content'** in line with the priority illegal harms we are particularly concerned about during a crisis (see paragraph 20.11). These are terrorism, threats, abuse and harassment (including hate), and foreign interference offence.

20.26 Similarly, our proposed definition of a 'crisis' in the Protection of Children User-to-User Code is as follows:

- A crisis is an extraordinary situation in which there is a serious threat to public safety to the United Kingdom either:
  - as a result of a significant increase in **relevant content harmful to children** on the service; and/or
  - which has caused or is highly likely to cause a significant increase in **relevant content harmful to children** on the service.

20.27 We have defined **'relevant content harmful to children'** in line with the types of priority content harmful to children that we are particularly concerned about during a crisis (see paragraph 20.11). These are hate and abuse, and violent content.

20.28 These definitions differ slightly from one another because the measures in the Illegal Content User-to-User Codes are recommended for the purpose of compliance with the illegal content safety duties set out in section 10 of the Act.[667] Similarly, the measures in our Protection of Children User-to-User Code are recommended for the purpose of compliance with the children's safety duties in section 12 of the Act.[668] In practice, many services will be subject to both measures[669] and should apply the two definitions together.

20.29 In developing this definition we have considered the following:

- The crises with which we are concerned are rare and unusual. We have therefore limited our definition to "extraordinary situations" which involve a "serious threat to public safety" to reflect the magnitude of the impact of such events on both online and offline settings. Nationwide riots, large scale terrorist attacks and/or inter-religious or inter-ethnic violence may, depending on the circumstances, satisfy our definition of a crisis.

- We also considered the definition of 'crisis' appearing in the EU's Digital Services Act 2024, which covers extraordinary circumstances leading to a serious threat to public security or public health in the EU or in significant parts of it.[670]

[667] See Illegal content Codes of Practice for user-to-user services, paragraph 2.7 [accessed 18 June 2025]

[668] See Protection of Children Code of Practice for user-to-user services, paragraph 1.2 [accessed 18 June 2025]

[669] This will be the case where a service is likely to be accessed by children (and is therefore subject to the children's safety duties in section 12 of the Act) and, as a result of the findings of its risk assessment, is within scope of both measures.

[670] Regulation (EU) 2022/2065 of the European Parliament and of the Council - EN - DSA - EUR-Lex. [accessed 5 May 2025].

- We consider that crises can occur as a result of a significant increase in illegal and/or content harmful to children.
- However, they can also cause a significant increase in illegal and/or content harmful to children, such as in the case of the Southport attack in 2024 where offline violence instigated an increase in anti-muslim and anti-migrant hate online.[671] Therefore, our definition covers both scenarios.

Providers should prepare and apply an internal protocol for responding to a crisis, including addressing the risk of an increase in illegal and/or content harmful to children on their service(s) during a crisis

20.30 We propose that providers should prepare and apply a written internal protocol for identifying and responding to a crisis as defined in paragraphs 20.24 and 20.26, including addressing the risk of an increase of illegal and/or content harmful to children on their services during a crisis and (where relevant) mitigating and managing the risk of the service being used for the commission or facilitation of a priority offence.

20.31 The crisis response protocol should include:

- Indicators identified by the provider that it will regularly monitor to determine whether a crisis is occurring or is likely to occur.
- How the provider will monitor those indicators.
- How the provider will keep the indicators under regular review to ensure that they remain relevant and that new indicators are identified as required. Where the provider determines that any of the indicators no longer remain relevant, or where new or updated indicators are identified, it should update its crisis response protocol accordingly.
- Details of a crisis response team, made up of representatives from all relevant internal teams including individuals of sufficient seniority to facilitate timely decision-making and action, that the provider will deploy in the event that it identifies a crisis is occurring or is likely to occur.
- How the provider will deploy the crisis response team.
- Systems and/or processes identified by the provider to address the risk of an increase in relevant illegal content and/or content harmful to children on the service, and (where relevant) the service being used for the commission or facilitation of a priority offence, during a crisis.
- How the provider will deploy the systems and/or processes.

[671] Institute of Strategic Dialogue and CASM Technology, 2024. Evidencing a rise in anti-Muslim and anti-migrant online hate following the Southport attack. [accessed 5 May 2025].

20.32 This element of the measure is aligned with current industry practices to some extent. There is evidence to show that some service providers already have some form of crisis response mechanisms in place.[672] [673] [674]

Indicators for monitoring and identifying a crisis

20.33 We propose that providers should identify indicators relevant to their services which signal that a crisis is occurring or is likely to occur.

20.34 Relevant examples of indicators may include, but are not limited to:

- information from law enforcement;
- information obtained via complaints processes, such as increases in volume of complaints;
- information obtained via content moderation processes, such as an increase in violative content as outlined in the service's terms and conditions, or increased virality of relevant illegal content and/or content harmful to children; and
- information from trusted flaggers.[675]

Systems and processes to address the risk of an increase in relevant illegal and/or content harmful to children

20.35 We propose that the crisis response protocol should include systems and/or processes that will be deployed to address the risk of an increase in relevant illegal content and/or content harmful to children, and (where relevant) the risk of the service being used for the commission or facilitation of a priority offence, during a crisis.

20.36 Relevant examples of such systems and processes may include, but are not limited to:

- Reallocating content moderation resources to meet the increase in illegal content and/or content harmful to children.
  - This example reflects the existing resourcing measure in the Illegal Content User-to-User Codes[676] and the Protection of Children User-to-User Code.[677] The respective measures recommend that in-scope providers resource their content moderation functions to give effect to their performance targets, having regard to (among other things) the propensity for external events to lead to an increase in demand for content moderation on the services.
  - We anticipate that service providers will build flexibility into the resourcing of their content moderation function, so that where an external event on a service causes a change in the levels of demand for content moderation, such that identified content

[672] Council of Europe, 2023.
[673] Meta, 2023. Crisis Policy Protocol. [accessed 5 May 2025].
[674] UK Parliament, 2025. 25 February 2025 – Social media, misinformation and harmful algorithms – Oral evidence. [accessed 5 May 2025].
[675] A trusted flagger is defined in the Illegal Content User-to-User Codes as: an entity which is a 'recommended trusted flagger' (this is defined further but is not relevant for present purposes) and any other person: a) whom the provider has reasonably determined has expertise in a particular illegal harm or harms; and b) for whom the provider has established a dedicated reporting channel. A trusted flagger is defined in the Children's Safety Code as: any entity for which the provider has established a separate process for the purposes of enabling the reporting of content which may include content harmful to children, based on the entity's experience.
[676] See ICU C6 of our Illegal content Codes of Practice for user-to-user services.
[677] See PCU C6 of our Protection of Children Code of Practice for user-to-user services.

across the system is dealt with efficiently. However, the recommended measures leave providers with the ability to update performance targets where appropriate. For more information on the existing measures, refer to the Illegal Content User-to-User Codes[678] and the Protection of Children User-to-User Code.[679]

- Increasing human content moderation resources.
- Adapting content moderation policies if they do not effectively capture the type of content emerging from the crisis.[680]
- Proactively identifying relevant illegal and/or content harmful to children stemming from the crisis, either through:
    - proactive sweeps (a practice through which human reviewers proactively review trending/viral content stemming from the crisis);
    - keyword searching; or
    - other content moderation technology.

20.37 At the onset of a crisis, providers of large services should set up a dedicated communication channel allowing law enforcement to contact them on matters related to the crisis.

Providers should conduct a post-crisis analysis

20.38 We propose that once a crisis is over (or after 90 days have passed, if sooner), providers should:

- conduct and record a post-crisis analysis assessing whether the crisis response protocol remains appropriate for addressing the risk of an increase in relevant illegal and/or content harmful to children on the service, and (where relevant) mitigating and managing the risk of the service being used for the commission or facilitation of a priority offence, during a crisis;
- keep a written record of these assessments;
- use the post-crisis analysis to make any changes to their service, for example its terms of service, any proactive technology that the service uses, or its content moderation processes.

20.39 We consider the timescale proposed is a reasonable period for providers to put in place more stable systems and processes once the initial demand associated with the crisis has lessened.

20.40 We are not proposing to recommend that services submit the post-crisis analysis to Ofcom or publish it. However, we may request the analysis and its findings in the future if necessary (for example, when exercising our regulatory supervision and/or enforcement functions).

[678] Ofcom, 2024.

[679] Ofcom, 2025.

[680] Note: Illegal Harms content moderation measure ICU C3 recommends that providers should already have processes in place to update internal content policies in response to evidence of new and increasing illegal harm on the services (as tracked in accordance with Measure ICU A4 in governance). ICU A4 recommends that providers should have an internal monitoring and assurance function to provide assurance that measures taken to mitigate risks of harms to individuals are effective on an ongoing basis. Therefore, providers should already have the processes in place to be able to update policies in response to a crisis.

Benefits and effectiveness at addressing risks

20.41 The proposed crisis response measures are intended to ensure that service providers can act promptly and effectively in reducing the spread of relevant content on their services, and (where relevant) mitigate and manage the risk that the service will be used for the commission or facilitation of a priority offence, by mitigating risks at each stage of an unfolding crisis.

20.42 Recommending that providers have a crisis response protocol will help ensure that they have contingency plans for managing a crisis, enabling them to respond rapidly to new developments. Without a plan in place for identifying and responding to crises, valuable time may be lost while providers attempt to establish whether a crisis has been initiated, assemble relevant personnel, and develop their response in an ad-hoc and reactive manner.

20.43 In addition, there are several benefits to service providers of having effective crisis response in place, such as those relating to reputation and user experience, which may act as a signal of stability to investors and/or advertisers.

20.44 The clear definitions and examples of crises that we have given in paragraphs 20.24, 20.26 and 20.29 will help providers identify a crisis when it occurs, and activate their crisis response protocol quickly.

20.45 Providers with mechanisms in place for monitoring indicators of a crisis will be better placed to identify the beginning of a crisis or an ongoing crisis and will be able to react swiftly by putting in place systems and/or processes to protect users from harm. Such indicators will also be useful for providers to monitor their overall operational efficiency and the resilience of their systems and processes when operating under high-pressure environments.

20.46 The proposed recommendation that service providers deploy a temporary team from all relevant parts of the business in the event of a crisis will improve the speed at which issues across the service are identified. It will also ensure that corresponding solutions are applied across the whole service (rather than in an ad hoc manner to parts of the service).

20.47 Deploying a crisis response team will safeguard against the risk of breakdowns in communication and enhance coordination between relevant teams in the high pressure setting of a crisis. This should improve the provider's ability to rapidly identify and remove illegal content and/or content harmful to children on the service.

20.48 We consider that large service providers having a dedicated communication channel available to law enforcement during a crisis would improve the speed and reliability of information exchange during crises. This would enable faster risk mitigation from providers and supporting coordinated public safety efforts at scale.

20.49 Lastly, we also consider that crisis response protocols should be improved over time. This is why we are also recommending for service providers to carry out a post-crisis analysis in the aftermath of a crisis.

20.50 A post-crisis analysis should drive improvement in providers' systems and processes for dealing with a crisis and identify gaps within the provider's wider trust and safety systems and processes. This will also mean we can formally request this report if necessary for the performance of our regulatory duties.

20.51 For these reasons, we provisionally consider that the proposed measures – when implemented alongside the existing measures set out in the Illegal Content User-to-User Codes and the Protection of Children User-to-User Code – would improve services' response times to crises and deliver important benefits given the heightened risk of harm during a crisis.

Impacts and Costs

20.52 In the following paragraphs we consider the main costs that service providers may incur as a result of our proposed measures (although we acknowledge that there may be other potential costs). The overall costs to service providers for implementing our proposed measures are likely to depend on a number of factors such as their existing systems and processes, number of users, technical complexity, and risk for the relevant harms.

20.53 Service providers will differ in their set-ups and risks for the relevant harms and the proposed measures allows for flexibility to service providers, which will enable services to implement policies which are appropriate, accurate and proportionate to their service.

Preparing and applying a crisis response protocol

20.54 Service providers will incur small one-off costs for preparing and applying a crisis response protocol. These will primarily consist of labour costs for the appropriate individuals to spend time developing and agreeing the required elements of the protocol (indicators to be monitored, details of the crisis response team, and systems and/or processes to address risks). We also recommend that large services should establish a dedicated communication channel by which they can be contacted by law enforcement. This would be undertaken (and any associated costs incurred) regardless of whether a crisis occurs on a service.

20.55 We assume that a service will monitor indicators arising out of existing systems and processes, and that this will not entail setting up new systems or processes.

20.56 We estimate that the work required to implement each of these elements will take between one day and one week of full-time work for the relevant individual or team, resulting in a total time spent between three days and four weeks to prepare and apply the crisis response protocol. This will vary depending on the service in question. For larger services, this team may comprise technical experts, as well as other professional occupation staff such as product managers, analysts and lawyers to support in developing and establishing a relevant policy and indicators. For large services, we also include one day of work for senior leader sign-off for each element of the protocol in our estimates. We estimate costs will range between £700 and £4,500 for a smaller service, with the smallest services being on the lower end of this estimate, and £4,600 and £11,300 for a large service.

Preparing and deploying a crisis response team

20.57 The size and composition of the crisis response team will vary depending on the provider. We anticipate that this will primarily be resourced from the provider's existing teams and that therefore the total incremental impact on service providers would be minimal.

20.58 There may be an opportunity cost for services associated with the reallocation of content moderators and other team members allocated to work on content stemming specifically from the crisis. However, we consider this proportionate to the risk being mitigated. Furthermore, the crisis response team will only be required in exceptional circumstances.

We assume service providers will adequately resource content moderation functions to meet their existing safety duties if they do not already do so.[681]

20.59 Those forming the crisis response team are also likely to require some form of training to familiarise themselves with the service's policies and processes in the event of a crisis.

20.60 For large services, we estimate the crisis response team would comprise around 5-10 full-time members of staff who would temporarily be moved from existing teams, plus one senior leader to oversee the team. This is likely to vary depending on the size and complexity of the service. This team would only be deployed in the event a crisis occurs and it is likely that many service providers would not need to call on this team each year.

20.61 We estimate training of the relevant staff will cost between £500 and £900 per annum for a small service and between £2,500 and £10,000 per annum for a large service (depending on the total size of the team), with the upper estimate reflecting a team of 10 members, assuming that each team member needs two days training.

20.62 As set out in our December 2024 Statement on Protecting People from Illegal Harms Online (December 2024 Statement) and our April 2025 Statement on Protecting Children from Harms Online (April 2025 Statement), the costs of implementing a content moderation system to review, assess and take down illegal content[682] or otherwise action content harmful to children[683] may be significant. We do not consider that the proposed crisis response measures require services to take a different approach to content moderation than that required by their safety duties.[684] We are of the view that any additional moderation costs incurred during the period of a crisis due to the proposed measures will be proportionate to the size and risk of a service and will be the minimum required to meet the existing duties listed.

Conducting a post-crisis analysis

20.63 For the post-crisis analysis element of the measure, we assume that the core crisis response team will spend one week analysing the events and the service provider's response. We estimate this will cost between £6,000 and £24,000 for large services and between £1,100 and £2,300 for smaller services. This cost will only be incurred in the rare event that a crisis affects the service.

[681] See ICU C6 of our Illegal content Codes of Practice for user-to-user services. [accessed 18 June 2025] This measure applies to providers of large user-to-user services and providers of multi risk user-to-user services. It is theoretically possible that a single risk user-to-user service would be in scope of the crisis response measures, but not ICU C6, however we are not aware of services which would fall into this category and expect this will be a very small number should any such services exist.

[682] See Chapter 2: Content Moderation of Volume 2: Service design and user choice in our December 2024 Illegal Harms Statement [accessed 18 June 2025]

[683] See Section 14: Content moderation for user-to-user services of Volume 4: What should services do to mitigate the risks of online harms to children in our April 2025 Statement on Protecting Children from Harms Online [accessed 18 June 2025]

[684] See ICU C1, C2, C5 and C6 of our Illegal content Codes of Practice for user-to-user services and PCU C1, C2, C5 and C6 of our Protection of Children Code of Practice for user-to-user services. [accessed 18 June 2025]

Wider market impact

20.64 We consider that these measures will encourage harm mitigation strategies and solutions to evolve based on an industry-wide standard of continual improvement through the post-crisis analysis.

20.65 Setting a regulatory standard for crisis response may incentivise third-party trust and safety providers to develop crisis response protocols and drive competition so that such services are more affordable for smaller providers.

20.66 Dealing with crises effectively will likely improve user experience and trust in online services. This may increase engagement and trust in services and act as a signal of stability to investors and advertisers.

20.67 As our proposed measures would impose a cost on in scope services regardless of size, there is a risk that these may disproportionately affect smaller services and act as a barrier to entry for new services and/or may cause existing services to exit the market. However, we provisionally consider that such risks and the overall cost burden for small services to be small and proportionate to their level of risk in a crisis. We therefore provisionally conclude that the risks of these measures having a significant impact on competition in the market are negligible.

Rights assessment

20.68 As acknowledged in our December 2024 Statement and our April 2025 Statement, content moderation is an area in which the steps taken by services may have a significant impact on the rights of individuals and entities – in particular, to freedom of expression under Article 10 of the European Convention on Human Rights (ECHR), freedom of association under Article 11, and privacy under Article 8. As these proposed measures are interconnected with, and enhance, existing content moderation measures, we have considered whether they are likely to result in interference with the rights of individuals and entities beyond the potential interference identified in connection with those existing measures, as set out in our December 2024 Statement and April 2025 Statement.[685]

Freedom of expression and association

20.69 The proposed measures are not prescriptive about how content is to be moderated during a crisis. Instead, they seek to secure that the provider has systems and processes in place by which it acts promptly and effectively to reduce the spread of illegal and/or content harmful to children on its services during a crisis. In general, the measure recommends that providers have a crisis response protocol in place, monitoring indicators of crises, deploying a crisis response team and identifying systems and process to address risks related to crises. It does not recommend that services remove content which is lawful. We consider that these measures do not have a direct negative impact on users' freedom of expression and association.

20.70 The proposed measures do not recommend that services have a higher tolerance for false positives in their content moderation processes during a crisis as compared with business

[685] See Volume 2: Service design and user choice, paragraphs 2.75-2.102 in our December 2024 Illegal Harms Statement [accessed 18 June 2025]

as usual. In other words, it does not recommend that they take down more content than they otherwise would have, nor does it recommend they adopt a different approach to identifying relevant illegal content or content harmful to children.

20.71 However, we recognise that a potential outcome of the proposed measures (depending on what providers put in their policy) is that providers may implement a crisis response protocol which prioritises speed of moderation over accuracy, which may result in a heightened risk of false positives (content being taken down from the services without it being illegal or harmful to children) and over-removal of content. Further, given the types of content the proposed measures relate to, there is a risk that false positives arise in relation to the most highly protected forms of speech such as religious expression (which could also impact users' rights to religion or belief under Article 9) or political speech, and in relation to the kinds of content that the Act seeks to protect such as content of democratic importance, journalistic content, and content from recognised news publishers.

20.72 Where providers adopt this approach, there is a potential for false positives to adversely impact users' rights under the ECHR, particularly Article 10. This potential interference is to be balanced against the very significant public interest in moderation of relevant illegal and/or relevant content harmful to children during a crisis. It is important to note that crises will occur infrequently, so the interference is likely to arise rarely. It is also important to note that during a crisis there is a serious threat to public safety in the UK, so the risk of harm deriving from relevant illegal and/or relevant content harmful to children may be very significant in that context. A higher risk of false positives may be more likely to be proportionate in these circumstances. However, we reiterate that this is not what these measures recommends – rather, it is a possible outcome of the measures.

20.73 There are also existing safeguards in place which minimise the risk of interference with the right to freedom of expression. Providers have incentives to limit the number of false positives that occur through content moderation, to meet user expectations and to minimise the costs of dealing with appeals. Existing measures on accuracy of decision making and on appeals[686] also act as a safeguard for freedom of expression. The Illegal Content Judgements Guidance (ICJG)[687] and the Protection of Children Harms Guidance[688] were prepared with careful consideration of rights of freedom of expression. Providers are encouraged to consult the ICJG and the Protection of Children Guidance when implementing the measures to assist with correctly identifying when freedom of expression considerations are particularly relevant to certain content.

20.74 Additionally, in accordance with the principles of the Act[689] and our duties under the Human Rights Act 1998,[690] we will have regard to the importance of freedom of expression and association, and the right to privacy, when making any decisions about enforcement in relation to these proposed measures. This acts as an additional safeguard for these rights.

20.75 The implementation of the proposed measures could also have positive impacts on the freedom of expression of users. The moderation of illegal or content harmful to children during a crisis could result in safer spaces online where users may feel more able to join

[686] See ICU C4, ICU D8 to D10 of our Illegal content Codes of Practice for user-to-user services and PCU C4, PCU D8 to D10 of our Protection of Children Code of Practice for user-to-user services [accessed 18 June 2025]

[687] See the Illegal Contents Judgement Guidance in our December 2024 Statement [accessed 18 June 2025]

[688] See the Guidance on Content Harmful to Children in our April 2025 Statement [accessed 18 June 2025]

[689] In particular, see section 1(3)(b) of the Act.

[690] Section 6 of the Human Rights Act 1998.

online communities and receive and impart information which is of particular use to them during a crisis.

20.76 In these circumstances – and noting that services have discretion about what to put in their crisis response protocols (which need not necessarily involve a higher tolerance for false positives) – we provisionally consider that any interference to users' rights to freedom of expression is proportionate. It is also be mitigated by the flexibility of the proposed measures, which allows providers to balance speed and accuracy when implementing crisis response protocols and is to be weighed against the very serious harm that can occur in crises.

Privacy and data protection

20.77 As the proposed measures are interconnected with (and enhance) existing content moderation measures, we have considered whether there are additional privacy and data protection impacts of the proposed measures to those identified in relation to those existing measures as set out in our December 2024 Statement.[691]

20.78 We consider the privacy and data protection impacts of the proposed measures to be inextricably linked and these are therefore assessed together. We do not expect the proposed measures to result in any significant additional interference with users' rights to privacy under Article 8 or their rights under data protection law. More importantly, providers processing users' personal data will still need to comply with applicable data protection legislation, including in relation to the accuracy of personal data. We consider these proposed measures to be compatible with data protection requirements.

20.79 Overall, and taking the benefits to users and affected persons into consideration, we provisionally consider that any impact on privacy and data protection rights from these proposals is proportionate.

Which providers should implement the measures

20.80 We propose that these measures will only apply to user-to-user services as current evidence shows that these are predominantly where the increase and spread of relevant illegal and/or relevant content harmful to children takes place during a crisis.

20.81 We are proposing that the set of measures should apply to:

- large user-to-user services that are at medium risk, and;
- user-to-user services of any size that are at high risk of any one of the following harms:
    - within priority illegal harms: terrorism, threats, abuse and harassment (including hate) and foreign interference;
    - within priority content harmful to children: abuse and hate, and violent content.

20.82 The measure we propose to include in our Protection of Children User-to-user Code of Practice will also be limited to services likely to be accessed by children.

[691] See our December 2024 Statement [accessed 18 June 2025]

20.83 For the reasons we have set out in this chapter, we provisionally consider that the benefits of applying these measures to large services which pose a medium risk of the relevant harms would be significant.

20.84 Whilst the costs of the measures may be more challenging for smaller service providers to initially absorb, we provisionally consider that it would be proportionate to apply the measure to high risk services regardless of their size. This is because the evidence we have assessed suggests that the dissemination of illegal and/or content harmful to children content often begins on smaller services before spreading to larger services.[692] Thus, smaller services can play a structural role in fostering perpetrator networks, disseminating illegal and/or content harmful to children, and catalysing offline violence in times of crises.

20.85 We provisionally consider that this approach is both proportionate and necessary given the evidence of the role of both large[693] [694] and smaller services[695] [696] in proliferating illegal and/or content harmful to children during crises.

Provisional conclusion

20.86 By having crisis response measures in place, service providers are more likely to act promptly and effectively respond to a crisis, including through the identification and removal of illegal and/or content harmful to children from their services. Ultimately, it is our intention that these proposed measures will safeguard UK users and communities – and by extension, our wider society – from the risks posed by online harms during a crisis.

20.87 We acknowledge that depending on how providers implement their crisis response policy, the measures may result in providers prioritising speed of moderation over accuracy. This may in turn involve a degree of interference with users' rights to freedom of expression and association. However, this would be a result of providers' choice and is not what our measure recommends. To the extent it does occur, we consider that such interference is likely to be proportionate given the circumstances of a crisis, where the risk deriving from illegal and/or content harmful to children may be very significant. Furthermore, we consider there are sufficient mitigations in place to safeguard those rights.

20.88 We are of the provisional view that the proposed measures are the least costly, least intrusive, and most effective measures to achieve the desired impact. Though the proposed measures would impose some costs on the relevant providers, we provisionally consider these costs to be proportionate given the high risk of harm being caused by illegal content and/or content harmful to children in a crisis and the impact the measures will play in mitigating those risks.

[692] Zheng, M., Sear, R., Illaria, L., Restrepo, N., Johnson, N., 2024. Adaptive link dynamics drive online hate networks and their mainstream influence. *npj Complex,* 1, 2. [accessed 5 May 2025].

[693] Rana, A., 2023. Content Moderation by Social Media Intermediaries – Examining Short Termism during the Ukraine Crisis, Harvard Law School, [accessed 5 May 2025].

[694] Institute for Strategic Dialogue, 2024.

[695] Casciani, D., BBC Verify, 2024. Violent Southport protests reveal organising tactics of the far-right, BBC, 2 August. [accessed 5 May 2025].

[696] Schwieter, C., 2022. Online crisis protocols – Expanding the regulatory toolbox to safeguard democracy during crises. [accessed 5 May 2025].

21. Broadening Appeals

Summary

We are making some amendments to the appeals measures in our Codes of Practice to help safeguard users' human rights. These amendments will extend the ability to appeal actions taken by service providers to cover action in relation to 'proxies' for illegal content and content harmful to children. In this context a proxy is a category of content which is prohibited by a provider's terms of service, which the provider is satisfied includes illegal content or content harmful to children.

This chapter describes these amendments.

Our Proposals

Number	Proposed measure	Who should implement this
Various ICU D measures[697]	Broadening the scope of appeals measures to ensure that they cover decisions taken on the basis that content was an 'illegal content proxy'	Various (as set out in the amended measures)
Various PCU D measures[698]	Broadening the scope of appeals measures to ensure that they cover decisions taken on the basis that content was a 'content that is harmful to children proxy'	Various (as set out in the amended measures)

Consultation questions

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

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

Introduction

21.1 Our Illegal Content User-to-user Codes of Practice and our Protection of Children User-to-user Code of Practice both contain measures related to 'relevant complaints'. For example, both ICU D1 and PCU D1 recommend that providers have systems and processes which enable prospective complainants[699] to make each type of 'relevant complaint'.

21.2 'Relevant complaint' is defined to include the following (emphasis added):

[697] ICU D2, D3, D4, D5, D6, D7, D8, D9, D10, D11, D12 and D13.

[698] PCU D2, D3, D4, D5, D6, D7, D8, D9, D10, D13 and D14.

[699] 'Prospective complaints' is defined in each of our user-to-user codes as 'United Kingdom users' and 'affected persons'. 'Affected person' is defined in section 20 of the Act.

Table 21.1: Definition of 'relevant complaint' in the User-to-user Codes

Illegal Content User-to-user Codes	Protection of Children User-to-user Code
Sub-para (c): complaints by a UK user who has generated, uploaded or shared content on a service if that content is taken down on the basis that it is illegal content.	Sub-para (c): complaints by a UK user who has generated, uploaded or shared content on a service if that content is taken down on the basis that it is content harmful to children.
Sub-para (d): complaints by a UK user if the provider has given a warning to the user, suspended or banned the user from using the service, or in any other way restricted the user's ability to use the service, as a result of content generated, uploaded or shared by the user which the provider considers to be illegal content.	Sub-para (d): complaints by a UK user if the provider has given a warning to the user, suspended or banned the user from using the service, or in any other way restricted the user's ability to use the service, as a result of content generated, uploaded or shared by the user which the provider considers to be content that is harmful to children.

21.3 In our User-to-user Codes we refer to these types of complaints as 'appeals'.[700] (As a consequence of introducing the defined term 'age assessment appeal' (explained in Chapter 18), we are also proposing to change the existing defined term 'appeal' to 'content appeal'. This is necessary to distinguish it from an 'age assessment appeal'. For the purpose of this chapter we continue to refer to 'appeals'.

21.4 Importantly, appeals do not extend to action that a provider takes on the basis that content was an 'illegal content proxy'[701] or a 'content that is harmful to children proxy'.[702] These terms are also defined in our Codes. This is content that a provider determines to be in breach of its terms of service, in circumstances where these terms are cast broadly enough to necessarily cover illegal content and content harmful to children (Primary Priority Content (PPC), Priority Content (PC) and Non-designated Content (NDC)). In this chapter we refer to content that is an 'illegal content proxy' or 'content that is harmful to children proxy' as 'proxy content'.

21.5 The limited scope of the appeals measures means that where a provider takes certain decisions on the basis that the user generated, uploaded or shared proxy content, the provider does not need to give the user a mechanism to appeal. Those decisions are:

a) A decision to take the content down;

[700] In the Protection of Children User-to-User Codes we refer to these types of complaints as a 'content appeal' to distinguish it from an 'age assessment appeal'. To date, this distinction has not been necessary in the Illegal Content User-to-User Codes. However, we are now proposing to introduce measures into the Illegal Content User-to-User Codes about 'age assessment appeals'. Consequently, we are proposing to replace all existing references to 'appeals' with 'content appeals'. We explain this in chapter 18.

[701] Illegal content proxy: In the Illegal Content User-to-user Codes, we defined "illegal content proxy" as content that a provider determines to be in breach of its terms of service, where: the provider has reason to suspect that the content may be illegal content; and the provider is satisfied that its terms of service prohibit the type of illegal content which it had reason to suspect existed.

[702] Content that is harmful to children proxy : In Codes, we define "content that is harmful to children proxy" as primary priority content (PPC) proxy, priority content (PC) proxy or non-designated content (NDC) proxy". This is content that a provider determines to be in breach of its terms of service, where: a) the provider had reason to suspect that the content may be relevant PPC, PC and/or NDC; and b) the provider is satisfied that its terms of service prohibit the type of relevant priority content which it had reason to suspect existed.

b) A decision to give the user a warning;
c) A decision to suspend, ban or in any other way restrict the user from using the service.

21.6 This may have a negative impact on that user's human rights. For example, in circumstances where a provider decides that a user shared proxy content and took their content down, but the content was not in fact proxy content, the user would not have a mechanism to bring this to the provider's attention under our existing measures.

21.7 Further, as part of this consultation we are recommending that:

a) users who share, generate or upload Child Sexual Exploitation and Abuse (CSEA) content or CSEA content proxies are banned from the service and prevented from returning.
b) providers prepare and apply a sanctions policy in relation to UK users who generate, upload, or share illegal content, illegal content proxy, content harmful to children and/or content that is harmful to children proxy (including PPC proxy, PC proxy and NDC proxy);

21.8 To complement these measures and safeguard users' human rights we consider it important that users can appeal decisions that providers make on the basis that their content is proxy content. It is also important that measures about the action that providers should take after determining an appeal are broad enough to reflect this position.

Our proposals

21.9 We are proposing various amendments to both Codes to achieve this objective. These amendments fall into two broad categories.

Broadening the scope of appeals measures to cover appeals based on proxy content

21.10 First, we are proposing to broaden the definitions of 'relevant complaint' and 'appeal' so that they cover decisions made by providers on the basis that content is proxy content. As a result, under ICU D1 and PCU D1, providers should have systems and processes which enable users to appeal such decisions. This will flow through to several other measures which relate to appeals, namely:

a) ICU D2 and PCU D2 recommend that providers have easy to find, easy to access and easy to use systems and processes for relevant complaints (including appeals).
b) ICU D3 and PCU D3 recommend that certain providers give specified information prior to the submission of a relevant complaint (including appeals).
c) ICU D4 and PCU D4 recommend that certain providers acknowledge receipt of relevant complaints (including appeals) and give the complainant an indicative timeframe for deciding the complaint.
d) ICU D5 and PCU D5 recommend that certain providers send further information about how the complaint will be handled upon receipt of relevant complaints.
e) ICU D6 and PCU D6 recommend that certain providers should enable complainants to opt out from communications following a relevant complaint.
f) ICU D7 and PCU D7 recommend appropriate action for relevant complaints about content which may be illegal content and content harmful to children respectively.

g) ICU D8 and PCU D8 recommend that certain providers should take appropriate action (among other things):

   i) determine relevant complaints which are appeals;
   ii) monitor their performance against performance targets relating to the time it takes to determine appeals and the accuracy of decision making for appeals; and
   iii) resource itself so as to give effect to those targets.

h) ICU D9 and PCU D9 recommend that certain providers should determine relevant complaints which are content appeals promptly.

i) ICU D10 and PCU D10 recommend steps that providers should take if, in relation to a relevant complaint that is an appeal, the provider reverses a decision that content was illegal content or content that is harmful to children respectively.

j) ICU D11 which recommends that providers inform complainants of specified matters when they receive relevant complaints which are not appeals about the use of proactive technology.

k) ICU D12 and PCU D13 which recommend appropriate action for relevant complaints about non-compliance with certain relevant complaints, namely:

   i) Nominating a responsible individual or team to ensure that such complaints are directed to an appropriate individual or team to be processed; and
   ii) Handling relevant complaints in a way that protects UK users (including children in the case of PCU D13) and within timeframes that the provider has determined are appropriate.

l) ICU D13 and PCU D14 ('Exception: manifestly unfounded complaints') which recommend that where providers receive relevant complaints that are not appeals, they may disregard the complaint in certain specified circumstances.

21.11 As a result of our proposed amendments these measures will apply to appeals of decisions made by providers on the basis that content is proxy content.

Consequential amendments

21.12 Second, we propose a number of 'consequential' amendments to support this position. These are set out in the table below.

Table 21.2: Proposed amendments for measures ICU D8 and D10 and PCU D8 and D10

Existing measure	Summary	Proposed amendment
ICU D8 'Appropriate action for relevant complaints which are appeals – determination (large or multi-risk services) PCU D8 'Appropriate action for content appeals – determination (services that are large or multi-risk)	These measures recommend that providers should have regard to specified matters in determining what priority to give review of appeals. In the ICU D8, these matters include (in summary): a) the seriousness of the action taken against the user or in relation to the content as a result of the decision that the content was illegal content. b) whether the decision that the content was illegal content was made by content identification technology (and if so, various other matters are specified). c) the past error rate on the service in relation to illegal content judgments of the type concerned. Similar matters are specified in PCU D8.	For ICU D8 We propose to broaden references to 'illegal content' to include 'or an illegal content proxy'. We propose to broaden references to 'illegal content judgments' to include 'judgments that content is an illegal content proxy'. For PCU D8 We propose to broaden references to 'content that is harmful to children' to include 'or a content that is harmful to children proxy'.
ICU D10 'Appropriate action for relevant complaints which are content[703] appeals – action following determination' PCU D10 'Appropriate action for relevant complaints which are content[704] appeals – action following determination'	These measures recommend certain steps that providers should take when they reverse a decision that content was illegal content (in ICU D10) or content that is harmful to children (in PCU D10).	We propose to broaden the reference to 'illegal content' in ICU D10 to also include 'or an illegal content proxy'. We propose to broaden the reference to 'content that is harmful to children' in PCU D10 to include 'or a content that is harmful to children proxy'.

Effectiveness

21.13 These measures ensure that users can appeal when service providers decide to take action which may restrict a user's ability to access content and help to prevent users from being

[703] We are proposing to change 'appeals' to 'content appeals' in this heading to make it clear that these measures do not concern age assessment appeals.

[704] We are proposing to change 'appeals' to 'content appeals' in this heading to make it clear that these measures do not concern age assessment appeals.

subject to erroneous content moderation decisions. An appeals process is also beneficial to service providers as it can help them to identify and action content that is illegal content and/or content harmful to children more accurately in the future. We propose that these benefits should extend to decisions taken on proxy content to protect users against excessive moderation and preserve their rights to freedom of expression.

21.14 We consider appeals that are dealt with systematically and accurately can help protect rights and build a more effective online safety system (ICU D8 and PCU D8). We consider that by the inclusion of illegal content proxy and content that is harmful to children proxy when setting and monitoring performance targets, providers may be able to better plan, configure and refine their processes to meet their goals.

21.15 We consider it is important to protect complainants' and users' rights to freedom of expression by ensuring that incorrect decisions that content is illegal content proxy and/or content that is harmful to children proxy are corrected by the provider through appropriate action (ICU D10 and PCU D10). This helps build more effective online safety and may also encourage providers to perform a more thorough review in their initial judgements to avoid having to process appeals.

Legal framework

21.16 Our proposals are underpinned by a number of provisions in the Act.

Providers' duties regarding complaints and safety measures

21.17 First, under section 21(2)(a) of the Act, user-to-user services have a duty to operate a complaints procedure that allows for 'relevant kinds of complaint' to be made. Such complaints include:[705]

a) complaints by a user who has generated, uploaded or shared content on a service if that content is taken down on the basis that it is illegal content or content that is harmful to children; and

b) complaints by a user of a service if the provider has:

i) given a warning to the user,
ii) suspended or banned the user from using the service, or
iii) in any other way restricted the user's ability to use the service,

c) as a result of content generated, uploaded or shared by the user which the provider considers to be illegal content or content that is harmful to children.

21.18 In our view, where a provider has chosen to moderate by identifying proxy content (rather than making illegal content judgments or determining whether content is harmful to children), it is unlikely to comply with these duties unless it considers appeals about decisions taken on the basis that content is proxy content.[706]

21.19 Second, user-to-user services have a duty to take or use proportionate measures to:

a) prevent individuals from encountering priority illegal content by means of the service,

[705] Sections 21(4)(c)-(d) and 5(c)-(d).

[706] We also note that Category 1 services subject to additional duties regarding complaints which are relevant in this context. These are set out in section 72(6)-(8) of the Act.

b) effectively mitigate and manage the risk of the service being used for the commission or facilitation of a priority offence, as identified in the most recent illegal content risk assessment of the service, and

c) effectively mitigate and manage the risks of harm to individuals, as identified in the most recent illegal content risk assessment of the service.

21.20 Similarly, user-to-user services that are likely to be accessed by children have a duty to take or use proportionate measures relating to the design or operation of the service to effectively:

a) mitigate and manage the risks of harm to children in different age groups, as identified in the most recent children's risk assessment of the service; and

b) mitigate the impact of harm to children in different age groups presented by content that is harmful to children present on the service.

21.21 In our view, a provider's decision to moderate by identifying proxy content (rather than making illegal content judgments or determining whether content is harmful to children) is unlikely to be proportionate unless it also allows users to appeal decisions made based on proxy content. This is because of the negative impact on users if they do not have a mechanism to appeal such decisions. For example, it is unlikely to be proportionate to ban a user from a service if they are unable to appeal the decision on which the ban is based. In this context, we note that Act's objectives are not limited to ensuring that regulated services are safe by design; they also include ensuring that such services are designed and operated in such a way that users' rights are protected.[707]

Ofcom's duties regarding measures recommended in Codes of Practice

21.22 We have a duty to prepare and issue one or more Codes of Practice for Part 3 services describing measures recommended for the purpose of compliance with certain duties, including the duties set out above.[708] We are therefore recommending these amendments for the purpose of user-to-user services' compliance with those duties.

21.23 We also have a duty to design measures in Codes of Practice in the light of certain principles, including the importance of protecting the right of users to freedom of expression within the law.[709] In our view there is a risk that some of our measures may disproportionately interfere with users' right to freedom of expression unless they are complemented by these proposed amendments. These proposed amendments help to safeguard users' right to freedom of expression. This is particularly so in the context of certain measures we are proposing within this consultation, namely our measures proposing that: a) users who generate, upload or share CSEA content or CSEA content proxies are banned from the service and prevented from returning, and b) providers should set and enforce a user sanction policy in respect to UK users who generate, upload or share illegal content and/or content harmful to children.

[707] Section 1(3) of the Act.

[708] Section 41(1)-(3) of the Act.

[709] Schedule 4 to the Act, paragraph 10(1)-(2)(a).

Impacts and costs

21.24 As we set out in our December 2024 Statement on Protecting People from Illegal Harms Online (December 2024 Statement) and our April 2025 Statement on Protecting Children from Harms Online (April 2025 Statement), service providers cannot comply with their duties under the Act without a robust appeals function and will therefore incur the costs of implementing existing appeals measures largely as a result of requirements of the Act. We consider that most service providers will have implemented changes that will deal with appeals centrally and will therefore not need to set up additional systems or processes as a result of our proposed amendments. We consider services are likely to use existing systems and processes to deal with these amendments given this will likely be the most practical and cost-effective approach. These overlaps mean additional costs associated with this measure are likely to be small. Furthermore, as set out in Paragraph 21.18 above, where a provider has chosen to moderate by identifying proxy content (rather than making illegal content judgments or determining whether content is harmful to children), it is unlikely to comply with its duties unless it considers appeals about decisions taken on the basis that content is proxy content. Our proposed amendments reflect this. Any additional costs arising are therefore a direct requirement of the Act.

Rights assessment

21.25 This section considers the impact of our proposed measures on users' rights to freedom of expression, to associate with others and to respect for private and family life. As explained in Chapter 3, restrictions on those rights must be necessary and proportionate.

Freedom of Expression and Association

21.26 As noted above, we consider that these amendments will safeguard users' rights to freedom of expression. In the absence of these amendments, there is a risk that certain other measures that we have recommended may disproportionately interfere with users' right to freedom of expression. This may occur where (for example) a provider makes an incorrect decision that content is proxy content and as a result takes one of the following actions without the user having an ability to appeal the decision:

a) Takes the user's content down, or takes other moderation action, in accordance with content moderation measures;

b) Bans users who have shared, generated or uploaded a proxy for CSEA content in accordance with our CSEA banning measure (see chapter 16: CSEA user banning);

c) Sanctions the user in accordance with our user sanctions measure (see chapter 17: User sanction policy).

21.27 The proposed amendments will ensure that such users are able to appeal such decisions. If their appeal were upheld the provider should reverse the action taken against the user or content (or both) as a result of that decision.[710] The existence of the appeal mechanism therefore safeguards users' right to freedom of expression.

Privacy and data protection

[710] ICU D10 and PCU D10.

21.28 All appeals systems and processes will involve the processing of personal data of individuals. It will therefore affect users' rights to privacy and their rights under data protection law.

21.29 We are proposing various amendments to existing measures concerning appeals. We assessed the privacy impacts of those measures in our December 2024 Statement[711] and April 2025 Statement.[712] To the extent there is an additional impact on the privacy of users because of these amendments, we consider this is likely to be proportionate to the benefits that these amendments provide to those users, including the protection they provider for users' rights related to privacy and data protection.

21.30 Furthermore, as set out above, the duties for a service provider to operate complaints systems and processes that enable relevant complaints are a requirement of the Act, including appeals. In our view, where a provider has chosen to moderate by identifying proxy content (rather than making illegal content judgments or determining whether content is harmful to children), it is unlikely to comply with these duties unless it considers appeals about decisions taken on the basis that content is proxy content. On this basis, we consider that that this measure is likely to constitute the minimum degree of interference required to secure that service providers fulfil their illegal content safety duties and child safety duties under the Act. In all circumstances providers should comply with their obligations under data protection law in relation to that processing.

Which providers should implement this measure

21.31 We are proposing various amendments to existing measures in our User-to-user Codes. Each of those existing measures applies to a specified class of providers. We are not proposing to change the class of providers to whom these measures apply.

Provisional conclusion

21.32 We consider appeals to be an important means of protecting complainants and users against excessive takedowns of content, and in preserving their right to freedom of expression. Determining appeals is beneficial to users as, in most cases, it enables content that has been incorrectly taken down to be restored and, where relevant, action taken against a user reversed. An appeals process is also beneficial to services as it can help them to identify and remove proxy content in the future more accurately. Where service providers do incur additional costs as a result of these proposed amendments, we consider these will be proportionate given the benefits.

[711] December 2024 Statement. Volume 2 Service Design and User Choice. Chapter 6. [accessed 13 June 2025]
[712] April 2025 Statement. Volume 4 What should services do to mitigate the risks of online harms to children. Chapter 16. [accessed 13 June 2025]

22. Combined Impact Assessment

Summary

In the preceding sections of this consultation document, we have assessed the individual impact of each of the proposed measures in the Codes. In this section, we assess the combined impact of the proposed measures as a package. Having considered the combined impact on different groups of services, we provisionally consider the package of measures to be proportionate.

Consultation questions

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

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

Introduction

22.1 In the preceding sections, we have assessed the impacts of each of the proposed measures individually and provisionally concluded that each is proportionate. In this chapter, we consider the combined impact of the proposed measures overall by looking at:

- Whether each measure has distinct benefits that contribute to how the overall package reduces risks of harms. This informs our views on whether the combined benefit of the package of measures may be significantly less than indicated when considering measures individually. This includes considering the impact of 'existing measures', by which we mean those recommended in both our December 2024 Statement on Protecting People from Illegal Harms Online (December 2024 Statement) and our April 2025 Statement on Protecting Children from Harms Online (April 2025 Statement). For example, if two measures targeted the same harm and one was very effective at reducing it, it could be disproportionate to impose both measures. We therefore consider the extent to which the benefits of different measures overlap.
- Whether the overall impact on services is proportionate, particularly for smaller services. In assessing this, we recognise that the costs of the whole package may be significant for some service providers, even where the cost of individual measures may not be significant.

22.2 We present the combined impact of the proposed measures for user-to-user services in scope of the Illegal Content Code, followed by user-to-ser services in scope of the Protection of Children Code.[713]

We consider the overall package proportionate for user-to-user services in scope of the Illegal Content Code

Each proposed measure delivers distinct and significant benefits

22.3 In this consultation we propose additional measures for user-to-user services in scope of the Illegal Content Codes to further improve online safety beyond existing measures recommended in the December 2024 Statement. For the list of proposed measures see Tables 21.1 and 21.2 later in this chapter.

22.4 Our view is that the benefits of these proposed measures do not substantially overlap with each other or with existing measures. Most of the proposed measures target specific severe harms and/or functionalities with clear link to harm, which our assessment shows are not adequately addressed by existing measures. For example, some target harms or functionalities that we have not specifically addressed in existing measures., such as intimate image abuse or one-to-many livestreaming, respectively.

22.5 Even where more than one proposed measure targets the same harm, they focus on different aspects of it and deliver distinct benefits, for example:

- Individual measures address the same harm in different ways. For example, both proactive technology and user banning measures target the distribution of child sexual exploitation and abuse (CSEA) material. Proactive technology is key to detecting and then removing child sexual abuse material (CSAM), particularly CSAM that had not previously been identified. User banning aims to remove access to services from users who share this material. We consider that together these measures will help reduce current CSAM content and reduce its future upload. As CSEA harm can manifest in different ways, more than one measure is required to address it.
- Certain measures in this consultation strengthen those recommended in the Illegal Content Codes. For example, we previously recommended that all providers of user-to-user services have a content moderation function designed to review, assess and take swift action against illegal content. Given the prevalence and severity of illegal content that is generated, uploaded and shared on user-to-user services, we propose in this consultation that, where accurate and effective proactive technology is available, services in scope of this measure should deploy such technology for the detection of relevant illegal content.

[713] For search services, we only propose one measure that applies to providers of large general search services in scope of the illegal harms code. While the overall cost – in additional to existing measures - could be significant, we consider it proportionate as it targets a harm that is not specifically addressed by existing measures. In addition, we expect large general search services to have the ability to implement this additional measure and we understand that some of these service providers already use hash matching to detect illegal content and are therefore likely to already have incurred most of the relevant costs of our proposed measure.

22.6 Overall, our assessment shows that each proposed measure has distinct benefits and removing it would diminish the overall effectiveness of the relevant package of measures. The measures proposed in this consultation build on existing measures and, in many cases, they are complementary. There is not significant duplication between the measures we are proposing.

We propose very few additional measures for all user-to-user services, including smaller low risk services

22.7 Recommending multiple costly measures for smaller services that are low risk is unlikely to be proportionate to the benefits they could provide in terms of reducing relevant harms. We are therefore proposing very limited additional measures for such services as shown in Table 21.1.

Table 21.1: Proposed additional measures for all user-to-user services in scope of the Illegal Content Code

Measure(s)	Short description	Who this measure applies to
User sanctions	Recommendation that services put in place a sanctions policy in relation to UK users who generate, upload or share illegal content and/or illegal content proxy with the objective of preventing future dissemination of relevant illegal content.	All user-to-user services
User banning and preventing return following detection of CSEA	Recommendation that services should ban users who share, generate, or upload CSEA, and those who receive CSAM, and take steps to prevent their return to the service for the duration of the ban	All user-to-user services

22.8 As shown in Table 21.1, we propose only two additional measures for all user-to-user services in scope of the Illegal Content User-to-User Codes. The first one is user sanctions for users sharing illegal content. Alongside this, we propose user banning as a specific sanction for users sharing CSEA. As explained in chapters 15 user banning and preventing return following detection of child sexual exploitation and abuse (CSEA) content and 16 user sanctions, these measures aim to reduce the risk of future dissemination of illegal content and CSEA and we consider them proportionate for all user-to-user services when assessed individually.

22.9 As both measures target preventing users from sharing illegal content there will be some overlap in the set-up costs for both measures. Therefore, there are likely to be cost efficiencies from services implementing both measures.

22.10 We consider the combined cost of these proposed measures to be proportionate for smaller low risk services, even when this is added to the cost of other measures that apply to them. As explained in chapters 15 and 16, we expect costs to be relatively low for smaller low risk services and there is flexibility for services to apply these measures in a cost-effective way. The costs of these measures are likely to increase with service size, as well as the number or complexity of harms on the services.

The overall cost for smaller user-to-user services with specific risks and/or functionalities is proportionate

22.11 We propose some measures for providers of user-to-user services that are in scope of the Illegal Content Codes if they have specific risks and/or functionalities. These are summarised in Table 21.2 and largely apply to relevant user-to-user services regardless of size.[714] However, some proposed measures apply to smaller services under stricter conditions compared to large services.[715]

Table 21.2: Proposed additional measures for user-to-user services with relevant risks or functionalities in scope of the Illegal Content Codes

Measure(s)	Short description	Who this proposed measure applies to
Recommender systems	Providers should design and operate their recommender systems to ensure that content indicated as potentially certain kinds of priority illegal content (relevant illegal content) is excluded from the recommender feeds of users	Providers of user-to-user services that have a content recommender system and are: medium or high risk for one or more of: hate, terrorism, suicide or foreign interference offences (FIO)
Livestreaming	The provider should have a mechanism to enable users to report that a livestream contains content that depicts the risk of imminent physical harm	Providers of user-to-user services offering one-to-many livestreaming that meet either of the following: are medium or high risk for one or more of relevant harms[716]

[714] In addition to the measures in Table 21.2 we propose some amendments to existing measures in the Illegal Content user-to-user Code as outlined in the relevant chapters, we consider the incremental cost of these amendments to be relatively low and provisionally consider them proportionate.

[715] For example, some proposed measures apply to smaller services if they are high risk for relevant harm(s) and have more than 700,000 monthly active UK users, while they apply to large services at high or medium risk for relevant harm(s).

[716] The relevant harms are terrorism content, grooming or image-based CSAM, assisting or encouraging suicide, hate, or harassment, stalking, threats and abuse offences. This includes services that only provide livestreaming and those where it is an ancillary feature.

Measure(s)	Short description	Who this proposed measure applies to
	The provider should, as part of its content moderation function (see ICU C1), ensure that human moderators are available whenever users can livestream using the service	Providers of user-to-user services offering one-to-many livestreaming that meet either of the following: are medium or high risk for one or more of relevant harms[717]
	Recommendations designed to reduce the risk associated with 'one to many' livestreams: Restricting comments, likes, or gifts to children operating livestreams Viewers should be prevented from capturing and/or recording videos and images of children's livestreams, including via in-service functionalities and (where technically feasible) third-party services.	Services that: a) offer one-to-many livestreaming, and b) it is possible for children to access the service, or a part of it.
User-to-user safety settings [718]	We are recommending amendments to ICU F1 and F2. These proposed changes would mean that services in scope of ICU F1 and F2 should implement these measures using one of the following approaches: using highly effective age assurance, as set out in the Part 3 HEAA Guidance; or applying ICU F1 and F2 to all users of the service	Providers of user-to-user services that meet either of the following: high risk for grooming large services that are at least a medium risk for grooming
Crisis response	The provider should prepare and apply an internal crisis response protocol. The provider should conduct a post-crisis analysis.	Providers of user-to-user services that are high risk or large medium risk of one or more of: hate, terrorism, harassment and foreign interference

[717] The relevant harms are terrorism content, grooming or image-based CSAM, assisting or encouraging suicide, hate, animal cruelty, or harassment, stalking, threats and abuse offences. This includes services that only provide livestreaming and those where it is an ancillary feature.

[718] This measure is an amendment of the ICU F1 and F2 to increase the effectiveness of the measures set out in our December 2024 statement, therefore our proposals will apply to the same set of services.

Measure(s)	Short description	Who this proposed measure applies to
	The provider should implement a dedicated communication channel by which law enforcement can contact them on crisis-related matters during a crisis.	Providers of large user-to-user services that are: medium risk for one or more of: hate, terrorism, harassment and foreign interference
Highly effective age assurance in the Illegal Content User-to-User Codes	Defining highly effective age assurance for the purposes of the Illegal Content User-to-User Codes and setting out principles providers should have regard to when implementing highly effective age assurance Providers should allow for appeals of highly effective age assurance decisions and take appropriate action when these are upheld	Providers of user-to-user services likely to be accessed by children that: use highly effective age assurance to determine which UK users of the service are child users for the purpose of targeting measures recommended in the Illegal Content Codes at such users
Proactive technologies	Providers should assess whether proactive technology to detect or support the detection of relevant illegal content is available, is technically feasible to deploy on their service, and meets the proactive technology criteria. If so, they should deploy it.	Providers of user-to-user services that: • are (1) large and medium risk, or (2) with more than 700,000 monthly users and at high risk for at least one relevant harm[719] • are file-sharing and storage services at high risk for image-based CSAM, regardless of size • are high risk for grooming, regardless of size
	Providers should assess existing proactive technology that they are using to detect or support the detection of relevant illegal content against the proactive technology criteria and, if necessary, take steps to ensure the criteria are met.	

[719] The relevant harms are image-based CSAM, CSAM URLs, grooming, fraud and other financial services offences (fraud), encouraging or assisting suicide (suicide).

Measure(s)	Short description	Who this proposed measure applies to
Perceptual hash-matching for intimate image abuse	Recommendation that providers use perceptual hash-matching to detect image-based intimate image abuse content so it can be removed	Providers of user-to-user services which are high risk of intimate image abuse and either: whose principal purpose is the hosting or dissemination of pornographic content; or are file-sharing and file-storage services; or have more than 700,000 monthly active UK users. Providers of large user-to-user services that are medium risk for intimate image abuse.[720]
Hash matching for terror content	Providers use perceptual hash matching to detect terrorism content so it can be removed	Providers of the following user-to-user services that enable regulated user-generated content in the form of photographs, videos or visual images (whether or not combined with written material) to be generated, uploaded or shared: Large high or medium risk of terrorism content high risk of terrorism content and have more than 700,000 monthly active UK users file-sharing services at high risk of terrorism content

22.12 The overall package of existing measures for the services in scope of the additional proposed measures in Table 21.2 could entail substantial costs.[721]

22.13 The additional proposed measures in Table 21.2 are targeted at smaller service providers where we have strong evidence of harm and/or where the risk of displacement is significant. For example, we propose hash matching measures for all file-sharing services and livestreaming measures for relevant service providers with a livestreaming functionality. The services in scope -including micro and small businesses - are still likely to pose significant risk and a substantial harm to users that would not be addressed in the absence of these measures.

[720] As set out in Chapter 10 perceptual hash matching for intimate image abuse, this measure also applies to providers of all large general search services in scope of the Illegal Harms Code, and we consider the combined impact proportionate for such services.

[721] As outlined in our Statement on Protecting People from Illegal Harms Online Volume 2, Chapter 13.

22.14 Some smaller services may struggle to resource the proposed measures. However, even if this were to happen, we would still consider the proposed package proportionate. This is due to the severity of risks of harms on such services and the distinct benefits of each of the measures in reducing these risks. We consider the costs are justified by the significant incremental benefits that arise from the measures individually and collectively.

22.15 For example, the proposed recommender system and livestreaming measures may lead to significant costs for some services, however they target different functionalities where the known risks are significant, and we consider the incremental benefit to be significant. Our crisis response measures complement existing measures in the Illegal Content user-to-user Codes and Protection of Children user-to-user Codes and provide additional important protections to users in the event of a crisis occurring on a service. We provide an alternative option for the age assurance for grooming measure, to apply ICU F1 and F2, to all users. This allows services to apply that measure in a way that is most appropriate for their service.

22.16 There may be cost synergies between measures when implemented together. For example, for services in scope of both the proposed measures that recommend hash matching, for IIA and terror content, we consider there could be overlaps in the hash-matching software for both types of illegal content and therefore this could reduce the overall cost of implementing both measures. We consider these measures to be effective in reducing severe and distinct harms and therefore consider them to be proportionate for in-scope services individually and cumulatively, regardless of potential synergies between them.

22.17 Some of the measures proposed in this consultation are likely to increase the cost of implementing some existing measures. For example, some proposed measures (such as hash-matching, livestreaming, and proactive technologies) aim to improve the services ability to identify relevant illegal content, which could increase the cost of implementing existing content moderation measures. We consider this proportionate as the proposed measures target services with specific harms or functionalities where we consider the risk to be significant and because the number of human moderators should increase with size and risk of the service.

The overall cost for large user-to-user services is proportionate

22.18 We recognise that large service providers could incur significant cost to implement existing measures, and the proposed additional measures (Tables 21.1-21.2) will increase this further. However, we expect providers of large services to have the resources to undertake these measures. Even if some providers have more limited resources, the increase in overall cost would be proportionate given the reduction in risk which could impact a significant number of users on these services.

We consider the overall package proportionate for User-to-User services in scope of the Protection of Children Code

22.19 Some of the additional measures proposed for illegal harms (in Tables 1 and 2), have equivalent measures proposed for user-to-user services in scope of the Protection of

Children Codes. As summarised in Table 21.3, these include user sanctions including smaller low-risk services.[722] mainly target user-to-user services that have specific risks of harms and/or functionalities, in additional to the user sanctions measure for those that prohibit one or more kinds of primary priority content (PPC), priority content (PC) and/or non-designated content (NDC).

Table 21.3: Proposed additional measures for user-to-user services in scope of the Protection of Children Code

Measure(s)	Short description	Who this measure applies to
User sanctions	Recommendation that services put in place a sanctions policy in respect of UK users who generate, upload or share content harmful to children and/or harmful content proxy, with the objective of preventing future dissemination of relevant content that is harmful to children.	All user-to-user services, likely to be accessed by children which prohibit one or more specific kinds of relevant PPC, PC and/or NDC
Proactive Technologies	Providers should assess whether proactive technology to detect or support the detection of relevant content harmful to children is available, is technically feasible to deploy on their service, and meets the proactive technology criteria. If so, they should deploy it.	Providers of user-user-services that are likely to be accessed by children that are: (1) large and medium risk; or (2) with more than 700,000 monthly active UK users and high risk for one or more of the relevant harms.[723]
	Providers should assess existing proactive technology that they are using to detect or support the detection of relevant content harmful to children against the proactive technology criteria and, if necessary, take steps to ensure the criteria are met.	

[722] In addition to the measures in Table 21.3 we propose some amendments to existing measures in the Protection of Children user-to-user code as outlined in the relevant chapters, we consider the incremental cost of these amendments to be relatively low and provisionally consider them proportionate.

[723] Relevant Content Harmful to Children: PPC, which includes pornographic, suicide, self-harm and eating disorder content.

Measure(s)	Short description	Who this measure applies to
Crisis Response	The provider should prepare and apply an internal crisis response protocol. The provider should conduct a post-crisis analysis. The provider should implement a dedicated communication channel by which law enforcement can contact them on crisis-related matters during a crisis.	Providers of all user-to-user services likely to be accessed by children that are: high risk of priority content that is harmful to children (abuse and hate; violent content); or large and medium risk of priority content that is harmful to children (abuse and hate; violent content)

22.20 These proposed measures aim to improve child safety beyond existing measures recommended in our April 2025 Statement. They have distinct benefits that do not substantially overlap with each other or with existing measures for the same reasons set out earlier for the equivalent illegal harm measures (see paragraph 21.4-21.6).

22.21 Some services in scope of these proposed additional measures could incur significant overall cost, particularly in addition to existing measures. We consider the overall cost proportionate for the same reasons set out for the equivalent illegal content measures in paragraphs 21.14-21.16.

Provisional conclusion

22.22 Our assessment shows that each measure has distinct benefits and contributes to reducing the risks online of illegal harms and improving safety for children online. These benefits go beyond those of the other measures in this package and those of the measures recommended in our previous statements. While the overall cost could be significant for some services, our assessment for the purposes of this consultation is that it is proportionate given the risks of harm to users and the incremental benefit of each measure in the package from reducing these risks. For the reasons set out in this chapter, our provisional conclusion is that the overall package of proposed measures is proportionate.

23. Statutory Tests

Summary

In designing the Codes, the Online Safety Act 2023 (the Act) requires Ofcom to have regard to a number of principles and objectives, set out in Schedule 4 to the Act. The Communications Act 2003 (the 2003 Act) also places a number of duties on Ofcom in carrying out our functions.

In this section, we set out the matters to which we must have regard under the Act and the 2003 Act, and explain the reasons why we think the recommendations in the Codes meet them. We provide further information regarding Ofcom's duties relating to the preparation of the Codes in our Legal Framework (Annex 16), and Annex 1, in which we set out our Equality Impact Assessment and Welsh language assessments.

Background

23.1 In designing the Codes, the Act requires us to have regard to a number of principles and objectives, set out in Schedule 4 to the Act. The 2003 Act also places a number of duties on Ofcom in carrying out our functions, including requiring us to have regard to the risk of harm to citizens presented by content on regulated services.

23.2 In Chapters 4-10, we set out our proposed recommendations. An overview of these recommendations can be found in Chapter 1, and our combined impact assessment of the recommendations can be found in Chapter 22. The draft Codes measures themselves can be found in full in Annexes 7, 8 and 9. We provide further information regarding Ofcom's duties relating to the preparation of the Codes in Chapter 3.

23.3 We consider that our proposals meet the requirements set out in Schedule 4 to the Act and section 3 of the 2003 Act. In this section, we take each of the relevant 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

23.4 As required by section 3 of the 2003 Act, in making the recommendations in the Codes, Ofcom has had regard to the matters set out below and to the risk of harm to citizens presented by content on regulated services.

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.

23.5 We have set out in this consultation how recommended Codes measures will mitigate the risks of illegal harm to all uses, and the risks of harm to children, thereby furthering their interests as well as the interests of citizens in the UK more generally.

23.6 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 recommendations, including any potential impacts on the provision of services to users.

23.7 We have also considered the rights of users and other interested persons in our rights assessment for each measure, where we consider any impacts of the measure on users' rights (children's and adults' as relevant), including their rights to freedom of expression and privacy, as required by the Act.

Section 3(3): In performing their duties under subsection (1), Ofcom must have regard in all cases to (a) the principles under which regulatory activities should be transparent, accountable, proportionate, consistent and targeted only at cases in which action is needed, and (b) any other principles appearing to Ofcom to represent best regulatory practice.

23.8 In the interest of transparency, accountability and fairness (and as required by the Act)[724], we are consulting stakeholders on our proposals and our consultation includes impact assessments for each of the measures we propose to include in the Codes. We are setting out clearly the evidence and assumptions used to arrive at our proposals. In Chapter 3 we explain how we approach impact assessments.

23.9 Our impact assessments of measures consider effectiveness, costs, rights, and other relevant factors and explain why we consider the measures are proportionate to the benefits to children and adults as relevant. We consider the proportionality of the package of the measures as a whole in Chapter 22. See our impact assessment guidance for more information on how we approach impact assessments.[725]

23.10 Our proposed measures are informed by our assessment of the risk posed by illegal content to all users and the risks of harm to children. 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.

Section 3(2)(g): In carrying out our functions, Ofcom are required to secure 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.

23.11 The changes we are proposing to make to our Codes of Practice are intended to mitigate the risks to users from illegal content, and the risks to children from content harmful to them. For example:

- Our recommendation on proactive technologies is designed to reduce the amount of illegal content and content harmful to children which reaches UK users;
- Our proposals on livestreaming would protect children by limiting opportunities for harmful interactions in the livestreaming environment.

[724] Under section 3(3) of the 2003 Act, Ofcom must, in the performance of their duties under subsection (1), have regard to the principles under which regulatory activities should be transparent, accountable, proportionate, consistent and targeted(only at cases in which action is needed. Ofcom must also have regard to any other principles appearing to us to represent best regulatory practice. This includes the public law duty to act fairly.

[725] Ofcom, 2023. Ofcom's approach to impact assessment. [accessed 16 June 2025]

23.12 These proposals are informed by our own assessment of the risks of harm, as set out in our Illegal Harm sand Children's Register of Risks[726] [727] and Risk Profiles[728] [729] plus some additional sources as detailed in the chapters of this consultation. In each chapter we explain how the measure will be effective in reducing risk and harm.

23.13 In relation to matters to which section 3(2)(g) in the 2003 Act 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.

23.14 As noted above the Illegal Harms and Children's Register of Risks and Risk Profiles set out the risks of harm posed by content on regulated services. These risks, alongside findings from services' risk assessments, largely inform what measures will be appropriate for a service provider to address the risk of harm to citizens. The measures included in the Codes vary across services based on their risk and size. In each chapter we discuss the risk of harm that we are seeking to address and why we think our proposed measures will be effective.

23.15 The Guidance on Content Harmful to Children[730] sets out examples of content, or kinds of content~~,~~ that we consider to be, or consider not to be, primary priority content and priority content that is harmful to children. The guidance is intended to support providers of Part 3 services that are likely to be accessed by children in making judgements about whether content on their service is content that is harmful to children as defined in the Act.

(b) The need for a higher level of protection for children than for adults.

23.16 Our Codes already ensure a higher level of protection for children than for adults. To the extent we are proposing amendments to the Codes, some of the measures we are proposing which impact children apply to all services (or all services with particular functionalities). These reflect steps that we expect all services should take to comply with the children's safety duties, regardless of their size and the risks they pose to children. Services that pose significant risks to children should take additional steps.

23.17 In this consultation we are proposing to strengthen our existing grooming measures by recommending the use of highly effective age assurance; this will keep children safer by ensuring that grooming protections are applied to the right users. Our measures on livestreaming also include specific measures that protect children, recognising the particular risks they may encounter when broadcasting livestreams.

23.18 The Codes also include a small number of measures for large services only, as there is significant scope to reduce the risk of harm for the many UK children that use them, and these providers have greater capacity to implement more costly measures.

(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.

[726] Ofcom, 2024. Register of Risks [accessed 16 June 2025]
[727] Ofcom, 2025. Children's Register of Risks [accessed 16 June 2025]
[728] Ofcom, 2024. Risk Assessment Guidance and Risk Profiles [accessed 16 June 2025]
[729] Ofcom, 2025. Children's Risk Assessment Guidance and Children's Risk Profiles [accessed 16 June 2025]
[730] Ofcom, 2025. Guidance on Content Harmful to Children [accessed 16 June 2025]

23.19 Our proposals, and the explanation in this document of how the proposed measures work, aim to provide clarity and tangible steps that services can take to meet their duties in the Act.

23.20 We have issued various guidance in the past that will support service providers in complying with measures that we are recommending. For example, the Illegal Content Judgements Guidance[731] and Guidance on Content Harmful to Children. We are also proposing to make some minor changes to the Part 3 HEAA guidance[732] as part of this consultation and are proposing further guidance to support providers in implementing our proposed measure on the assessment and use of proactive technologies.

23.21 We have explained in section Chapter 3 and Annex 16 that the Act provide that services which choose to implement the measures in the Codes will be considered as complying with relevant duties. We have also explained that service providers may seek to comply with their safety duties by choosing to take alternative measures.

(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.

23.22 The Risk Assessment Guidance[733] and Children's Risk Assessment Guidance[734] take account of the nature and sizes of services, for example in recommending what evidence providers should take into consideration to support their risk assessments.

23.23 We have clearly identified in the proposed Code measures the types and sizes of services we recommend adopt each measure, for the reasons given in each chapter. The Act requires us to ensure measures are proportionate, and we recognise that the size, capacity, and risks of services differ widely. We therefore do not take a one-size-fits-all approach. Instead, we have set out what types of service we think should use specific safety measures to comply with their duties, with the most extensive expectations placed on the riskiest services.

(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.

23.24 In this document we propose various measures regarding the use of such technologies – these include a new principles-based measure regarding the use of proactive technology to identify illegal content and content harmful to children, as well as recommending the use of hash matching technology to identify intimate image abuse and terror content.

23.25 Section 3(4) of the 2003 Act[735] 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.

[731] Ofcom, 2024. Illegal Content Judgements Guidance [accessed 16 June]
[732] Ofcom, 2025. Guidance on highly effective age assurance for Part 3 services [accessed 16 June]
[733] Ofcom, 2024. Risk Assessment Guidance and Risk Profiles [accessed 16 June 2025]
[734] Ofcom, 2025. Children's Risk Assessment Guidance and Children's Risk Profiles [accessed 16 June 2025]
[735] As amended by section 82 of the Act

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.

23.26 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. To the extent that we are proposing measures related to the illegal content safety duties, these aim to prevent crime and disorder. In relation to the opinions of consumers in relevant markets and of members of the public generally, we have taken account of the views of those with lived experience of harm (for example in relation to our CSEA banning measure) and stakeholder responses to our previous consultations. We will also take account of any opinions of consumers in relevant markets that we receive in response to this consultation.

23.27 In proposing measures in pursuit of children's safety duties, 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. In our equality impact assessments, we have considered the needs of persons of protected and listed characteristics. We have also considered our Welsh language obligations. See Annex 1.

Schedule 4, Online Safety Act 2023

23.28 As required by paragraph 1 of Schedule 4 to the Act, Ofcom has considered the appropriateness of provisions of the proposed additional measures for the Codes of Practice to different kinds and sizes of Part 3 services and to providers of differing sizes and capacities and we have set out our reasons for proposing some Codes recommendations to services of different kinds, sizes and capacities.[736]

23.29 We have had regard to the principles in Schedule 4 to the Act, 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.[737]

23.30 We have clearly identified which proposed measures apply to what types and sizes, for the reasons given in each relevant section of this consultation.

[736] See also section 3(4A)(d) of the 2003 Act.

[737] See also section 3(4A)(c) of the 2003 Act.

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.[738]

23.31 Having regard to the need for it to be clear to providers of regulated services how they may comply with their duties, we have aimed to be as clear as possible and to include an appropriate level of detail in this consultation and the supporting documents, consistent with acting proportionately. We have sought to be sufficiently detailed and precise while ensuring our proposals are technically feasible and proportionate for the wide range of services in scope of the Act. Our approach to the Codes strikes the balance between providing certainty about what providers need to do and allowing them flexibility to implement measures in a way that works in the context of their own services and is proportionate. For example, our proposed measure on user sanctions gives some flexibility to service providers in determining what sanctions should be applied.

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;

23.32 We have clearly identified in the draft Codes which measures apply to what types and sizes of services, for the reasons given in each chapter. We have considered proportionality and technical feasibility, where appropriate, as part of our assessment of impacts across this consultation, including in Chapter 22. For example, our principles-based recommendation on proactive technology makes it clear that we do not expect services to implement technology to detect particular types of content where this is not technically feasible.

23.33 We have taken 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 recommend. We consider effectiveness, costs, rights impacts, and other relevant factors in our assessment of proportionality.

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.

23.34 We have identified the relevant risks of harm that the measures address and explained why we consider each proposed measure is proportionate in the light of those harms. As required by section 3(4A)(b)(ii) of the 2003 Act, 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 the Illegal Harms and Children's Register of Risks. Where appropriate, we have clearly identified which measures would apply to what types and sizes of services, for the reasons given in each relevant section of this statement.

23.35 Overall, the Codes place more demanding expectations on services that pose greater risks. Having regard to the desirability of encouraging investment and innovation in the markets for regulated services and these technologies, our recommendations provide sufficient flexibility for services. Our impact assessment for each measure, as well as our combined

[738] See also section 3(4A)(c) of the 2003 Act.

impact assessment, also take into account the cost to services as we acknowledge additional costs can affect investment and innovation.

Ofcom's online safety objectives

User-to-user services

23.36 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 user-to-user services as set out in this sub-section.

23.37 All our proposals, in line with paragraph 11 of Schedule 4 to the Act, 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.

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.

23.38 The Codes already include measures related to governance and accountability, and we are not proposing to change these.

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.

23.39 In our Protecting People from Illegal Harms Online Statement[739] (December 2024 Statement) (paragraph 14.14) and our Protecting Children from Harms Online Statement[740] (April 2025 Statement) (paragraph 21.53), we said that the content moderation, automated content moderation and reporting and complaints measures in our existing codes were set having regard, amongst other things, to the number of users of the service and its user base. To the extent we are recommending changes to these measures, or additional such measures, we consider these 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.

23.40 Our existing codes recommend measures related to terms of service, we are not proposing to amend these or introduce additional such measures. We note that we are recommending additional measures involving proactive technology. In this regard, our existing measures related to terms of service ensure that providers include information about proactive technology in their terms of service.

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.

23.41 We are proposing a number of measures related to reporting and complaints. Our livestreaming measures seek to ensure it is easy for a UK user to report a livestream which

[739] Ofcom, 2024. Statement: protecting People from Illegal Harms Online [accessed 16 June 2025]

[740] Ofcom, 2025. Statement: Protecting Children from Harms Online [accessed 16 June 2025]

features a risk of imminent physical harm. We are also proposing to amend the appeals mechanisms in the Codes to ensure they function effectively.

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.

23.42 A number of existing measures are compatible with this objective. For example, user reporting and complaints measures ICU D2 and PCU D2 recommend that systems and processes should be easy to access and easy to find, helping child users report content harmful to children to the service provider. Many of our additional proposals that we are now consulting on are also compatible with this objective; for example, our proposal that service providers should use highly effective age assurance to target livestreaming protections at children.

Paragraph 4(a)(vii): a service should be designed and operated in such a way that the different needs of children at ages are taken into account.

23.43 In our December 2024 Statement at chapter 10: 'Terms of Service', and chapter 8: 'U2U Settings, Functionalities, and User Support', we set out how we had regard to the different needs of children at different ages. Further, in our April 2025 Statement, we noted that service providers have a duty, as part of their children's risk assessment, to assess their user base, including separate consideration to children in different age groups on the service and assessing how the design and use of the service affects the level of risk of harm to children and that we therefore consider that this objective will be secured in particular via the children's risk assessment duties and the Children's Risk Assessment Guidance.

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.

23.44 We are proposing to recommend that users who share CSEA material on a service are banned from the service, by having their access removed and being prevented from returning.[741] We are also proposing that user to user services have a policy regarding sanctions where UK users share other illegal content. We consider that these proposals are consistent with this objective.

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.

23.45 Our Protection of Children User-to-user Code already recommends the use of highly effective age assurance in certain circumstances. We are also proposing to recommend that highly effective age assurance is used by services for the purpose of complying with certain proposed recommendations in the Illegal Content User-to-user Codes.

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,

[741] Consistent with paragraph 11 of Schedule 4 to the Act, this measure applies to the design and operation of the service in the UK, and as it affects UK users.

- functionalities of the service, and
- other features relating to the operation of the service.

23.46 Our recommender systems measure proposes that content indicated as potentially priority illegal content is excluded from the recommender feeds of users. Our measures for livestreaming will aim to protect child users from grooming and child sexual abuse as well as protect all users from viewing illegal content where there is a risk of imminent physical harm.

23.47 We have not at this stage consulted on recommending measures relating to paragraph 4(a)(v) – "(in the case of a Category 1 service) users are offered features to increase their control over certain categories of content that 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.

Search services

23.48 We are proposing limited changes to our search Codes. As required by paragraph 3 of Schedule 4 to the Act, we have ensured that our proposed recommendations, in the context of our existing Codes, are compatible with the pursuit of the applicable online safety objectives as set out in this sub-section.

23.49 All the measures in the Codes, in line with paragraph 11 of Schedule 4 to the Act, 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.

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.

23.50 In our December 2024 Statement and our April 2025 Statement, we have set out recommended measures that, amongst other things, have regard to the kind and size of a search service.

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.

23.51 We are proposing that large search services should use hash matching for the purposes of identifying intimate image abuse content, reducing the risk that such content reaches a large number of UK users.

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 27 and 29.

23.52 In our December 2024 Statement and our April 2025 Statement, we have set out recommended measures that, amongst other things, are designed to ensure this statement is accessible to children.

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.

23.53 In our December 2024 Statement and our April 2025 Statement, we have set out recommended measures that support this goal including in relation to reporting and complaints.

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.

23.54 In our December 2024 Statement and our April 2025 Statement, we have set out recommended measures that support this goal. We note that the measures in the Protection of Children Search Code are designed to secure a higher level of protection for children than for adults.

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.

23.55 In our December 2024 Statement and our April 2025 Statement we set out recommended measures to support this goal including in the Reporting and Complaints and PAS chapters of our 2024 Statement and the draft Protection of Children Code for search services.

Paragraph 5(b): a service should be assessed to understand its use by, and impact on, children in different age groups.

23.56 As set out in our April 2025 Statement, service providers have a duty, as part of their children's risk assessment, 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.

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

23.57 In this consultation we are proposing to recommend that certain providers of search services implement hash-matching technology to detect intimate image abuse content. This will protect users of the service from harm (as well as other individuals who are not users of the service). We consider our recommendations in Volume 1, Chapter 5; Volume 2, Chapter 3; and Volume 2, Chapter 9 of our December 2024 Statement, and Section 11, Section 15 and Section 19 of our April 2025 Statement, to be compatible with this objective.

Schedule 4 requirements on Content of Codes of Practice

User-to-user services

23.58 Codes of practice that describe measures recommended for the purpose of compliance with duties set out in section 10(2) or 10(3) (illegal content) and section 12(2) or (3) (safety duties protecting children) of the Act, must include measures in each of the areas of a service listed in section 10(4) (illegal content) and section 12(8) (safety duties protecting children). These provisions apply to the extent that inclusion of the measures in question is consistent with:

- 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;
- 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
- 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.

23.59 In our December 2024 Statement (paragraph 14.18) and our April 2025 Statement (paragraph 21.94) we explained that our User-to-user Codes included recommendations in each of the areas of a service listed in section 10(4) and 12(8). To the extent we are making further recommendations they are in the following areas:

- Design of functionalities, algorithms and other features: proposals in relation to recommender systems and the user of automated tools to identify potentially illegal content and content potentially harmful to children.
- 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: measures relating to the removal of access for users who share CSEA material, and a wider user sanctions measure.
- Content moderation, including taking down content: our proactive technology measures will help support the detection of content for moderation; and our recommendation that services offering livestreaming functionalities have human moderators available.
- User support measures: improved reporting functions for livestreaming, measures to protect children when livestreaming and the use of HEAA to target existing measures to address grooming in the Illegal Content Codes of Practice.

23.60 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:

- 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.
- The importance of protecting the privacy of users.

Search services

23.61 Codes of Practice that describe measures recommended for the purpose of compliance with a duty set out in section 27(2) or (3) (illegal content) and section 29(2) or (3) (safety duties protecting children) of the Act, must include measures in each of the areas of a service listed in section 27(4) (illegal content) and section 29(4) (safety duties protecting children). This provision applies to the extent that inclusion of the measures in question is consistent with:

- 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;
- the principle that the measures described in the Code of Practice must be proportionate and technically feasible; and
- 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.

23.62 In our December 2024 statement (paragraph 14.23) and April 2025 statement (paragraph 21.98) we explained that our Search Codes included recommendations in each of the areas of a service listed in sections 27(4) for search services in the following areas of a service listed in section 27(4) and 29(4). To the extent we are making further recommendations they are in the following areas:

- Design of functionalities, algorithms and other features relating to the search engine: recommending the use of hash matching to identify intimate image abuse content

23.63 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:

- 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.
- The importance of protecting the privacy of users.

The overview section in this document is a simplified high-level summary only. The proposals we are consulting on and our reasoning are set out in the full document.