# Protecting people from illegal harms online

## Volume 2: Service design and user choice

**Statement**

Published 16 December 2024

# Contents

## Section

# 1. Service Design and User Choice (Volume 2) – Introduction

Our Codes work together to create an overall safer experience for users and cover three broad areas.[1]

In this volume, we set out and explain our decisions that contribute to ensuring online services are **designed and operated with safety in mind** and that there is improved **choice for users to enable more control over their online experiences**. Each measure set out in this volume contributes to one of these two strategic objectives.

## What we are trying to achieve

1.1 As set out previously in this Statement, we expect implementation of the Online Safety Act 2023 ('the Act') to ensure people in the UK are safer online by delivering four outcomes:

a) stronger safety governance in online services;
b) online services are designed and operated with safety in mind;
c) greater choice for users so they can have more meaningful control over their online experiences, and
d) greater transparency regarding the safety measures services use, and the action Ofcom is taking to improve them, to build trust.

1.2 In this Volume we set out and explain decisions relating to **ensuring online services are designed and operated with safety in mind**, and that there is improved **choice for users to enable more control over their online experiences**. By putting in place mitigations that recommend systems and processes that are proportionate to mitigate illegal harm, and that illegal content is taken down quickly, we are taking important steps in support of the second objective. By recommending measures that ensure all users have tools to manage their online experiences and that vulnerable groups, in particular, have safer experiences online, we are furthering the third objective.

## What decisions we have made towards these objectives

1.3 The Codes set out actions a service provider can take to mitigate the risk of illegal harm. These include for example, effectively training and resourcing content moderation teams, or making a service's complaints and reporting functions easy to find and easy to use.

[1] The three broad areas are: (1) stronger safety governance in online services (2) online services are designed and operated with safety in mind; and (3) greater choice for users so they can have more meaningful control over their online experiences. For more information on these, see our 'approach to Codes' chapter.

1.4 Our Illegal Content Codes of Practice for user-to-user ('U2U') and search services, will contribute to our strategic objectives in relation to **safer design** and **greater choice and control over online experiences.** The decisions we have taken in relation to safer design can broadly be broken down into: decisions designed to make it harder for people to post illegal content or engage in illegal conduct online in the first place; decisions designed to increase the probability that where people do post illegal content/engage in illegal conduct online service providers are able to detect and address this as quickly as possible; and decisions designed to make it less likely that illegal content goes viral in the period before it is detected and removed. This volume is divided up into the following chapters which describe and explain these decisions:

a) In Chapters 2 and 3 we describe and explain a series of decisions we have taken about how U2U and search service providers respectively should configure their **content and search moderation** functions. These decisions are intended to ensure that service providers' content moderation functions work effectively. We consider this will increase the probability that providers detect and address illegal content quickly and accurately.

b) Chapters 4 and 5 describe and explain a series of decisions we have taken about what **automated tools** providers of U2U and search services respectively should use **to detect potentially illegal content**. The aim of the measures we outline in these chapters is to ensure that illegal content is detected and addressed quickly and accurately. The decisions set out in these chapters primarily focus on the use of automated tools to detect CSAM. We intend to build on them next Spring by publishing a further consultation on additional automated content moderation tools service providers should use.

c) Chapter 6 describes and explains the decisions we have taken about how service providers should configure their **reporting and complaints functions**. The decisions outlined in this chapter make an important contribution to our strategic objective of ensuring services are designed with safety in mind. Where providers operate easy to use reporting and complaints functions and take appropriate action in response to complaints, they are more likely to be able to detect and address illegal content quickly and effectively.

d) Chapter 7 describes and explains decisions we have taken about how service providers should test their **recommender systems**. Our recommendation here will help service providers make more informed choices about the design of their recommender algorithms and are better placed to manage risks associated with their algorithms.

e) Chapter 8 describes and explains a number of decisions we have taken about **safety defaults** we expect service providers to put in place to protect children and **supportive information** they should provide to children. Our decisions are an important part of our work to ensure services are designed with safety in mind and are intended to make it harder for perpetrators to contact children with the intention of grooming them.

f) Chapter 9 describes and explains a number of decisions we have taken about the **design of search services**.

g) Chapter 10 describes our decisions about what service providers should do to ensure their **terms of service** and **publicly available statements** are clear and accessible. Consistent with our third strategic objective, the decisions in this chapter are intended to help users make more informed choices about which services they use.

h) Chapter 11 describes decisions we have taken into the circumstances in which providers should **prevent perpetrators from using their services**. The decisions in this chapter relate primarily to proscribed terrorist organisations. Next Spring we will be consulting on further proposals around user access and CSAM.

i) Chapter 12 sets out a series of decisions we have taken about **tools** services should offer users **to protect themselves from illegal content** and conduct online, consistent with our third strategic objective.

j) We assess the impact of each of our decisions in the relevant chapters above. Chapter 13 looks at all the decisions we have taken in the round and sets out why we think that **cumulatively their impact is proportionate**.

k) Chapter 14 outlines the different principles and objectives set out in **Schedule 4** to the Online Safety Act and **section 3** of the Communications Act and explain the reasons why we think our recommendations for our Illegal Content Codes of Practice meet these requirements.

# 2. Content Moderation

## What is this chapter about?

Content moderation is when a service provider reviews content to decide whether it is permitted on its service and takes appropriate action to handle it. Content moderation is used by providers to address a wide variety of illegal content as well as legal content that does not comply with a service's terms of service (which we call 'illegal content proxy'). This chapter sets out the content moderation measures we are recommending, why we are recommending them, and to which user-to-user (U2U) services they should apply.

## What decisions have we made?

We are recommending the following measures:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| ICU C1 | Providers should have systems and processes designed to **review and assess content** the provider has reason to suspect may be illegal content (part of its 'content moderation function'). | Providers of U2U services. |
| ICU C2 | Providers should have systems and processes designed to **swiftly take down illegal content and/or illegal content proxy** of which they are aware (part of their 'content moderation function'), unless it is currently not technically feasible for them to achieve this outcome. | Providers of U2U services. |
| ICU C3 | Providers should **set and record internal content policies**. | • Providers of large U2U services.<br>• Providers of multi-risk U2U services. |
| ICU C4 | Providers should **set and record performance targets** for their content moderation function. | • Providers of large U2U services.<br>• Providers of multi-risk U2U services. |
| ICU C5 | Providers should prepare and apply a policy in respect of the **prioritisation of content for review.** | • Providers of large U2U services.<br>• Providers of multi-risk U2U services. |
| ICU C6 | Providers should **resource their content moderation function**, so as to give effect to measure ICU C3 and measure ICU C4. | • Providers of large U2U services. |

| | | |
|---|---|---|
| | | - Providers of multi-risk U2U services. |
| **ICU C7** | Providers should ensure **individuals working in moderation** (non-volunteers) **receive training and materials** that enable them to **fulfil their role** in moderating content, including in relation to measure ICU C1, measure ICU C2 and measure ICU C3. | - Providers of large U2U services.<br>- Providers of multi-risk U2U services |
| **ICU C8** | Providers should ensure **volunteers** in their content moderation functions **have access to materials** that enable them to **fulfil their role** in moderating content, including in relation to measure ICU C1, measure ICU C2 and measure ICU C3. | - Providers of large U2U services.<br>- Providers of multi-risk U2U services. |

### Why have we made these decisions?

Effective content moderation systems are able to identify, and prioritise the swift removal of illegal content. Content moderation therefore plays a hugely important role in combatting illegal content. Providers with ineffective content moderation functions may face increased risk of harm on their services. Our analysis suggests that harm to users will be reduced where providers set content policies, resource and train their content moderation teams appropriately and take into account the likely severity of content and the risk the content will be encountered by a high number of UK users when deciding what potentially harmful content to prioritise for review. Given the diverse range of providers in scope of the new regulations, a one-size-fits-all approach to content moderation would not be appropriate. Instead of making very specific and prescriptive recommendations about content moderation, we have therefore decided to make a relatively high-level set of recommendations which would allow services considerable flexibility about how to set up their content moderation teams. We have focussed the most rigorous proposals in this area on services which are large or multi-risk. This will help ensure that the impact of the measures is proportionate. Similarly, the flexibility built into our proposals will make it easier for providers to carry them out in a way which is cost-effective and proportionate for them.

## Introduction

2.1 Content moderation is when a service provider reviews content to decide whether it is permitted on its service and takes appropriate action to handle it.[2] It is used by providers to address a wide variety of illegal harms as well as legal content that does not comply with their content policies. While content policies usually prohibit the posting of illegal content,

[2] Gillespie, T., and Aufderheide, P., 2020. Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. Internet Policy Review; Trust & Safety Professional Association [accessed 13 November 2024]. Singh, S., 2019. What Is Content Moderation? Everything in Moderation: An Analysis of How Internet Platforms are Using Artificial Intelligence to Moderate User Generated Content. [accessed 24 November 2024].

they do not necessarily closely reflect the requirements of any single legal system due to the global nature of many services.[3]

2.2 Content moderation systems and processes differ between services and are designed to meet specific needs and contexts. Content moderation can be carried out by humans, automated tools or a combination of the two.[4] We note that service providers use a combination of techniques to moderate content and that there are benefits and risks to differing moderation systems.

2.3 The measures within this chapter aim to secure that providers make appropriate decisions about suspected illegal content and take appropriate action to protect users.

## The Online Safety Act 2023

2.4 Under section 10 of the Online Safety Act 2023 ('the Act'), providers of regulated user-to-user (U2U) services have several duties.

a) They must take proportionate steps to 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 risk of harm to users as identified in the illegal content risk assessment (section 10(2)(a), (b) and (c)).
b) They must have proportionate systems and processes in place designed to minimise the length of time for which any priority illegal content is present (section 10(3)(a)).
c) They must have proportionate systems and processes in place designed to swiftly take down any illegal content when they are alerted by a person to its presence or they become aware of it in any other way (section 10(3)(b)).

2.5 As set out in chapter 6 of this Volume: 'Reporting and complaints', providers also have a duty to respond to complaints about illegal content and to handle appeals when action is taken on content or against users because the content is identified as illegal.

2.6 In practice, compliance with these duties would be very difficult without a process for determining whether or not content ought to have appropriate action taken on it, or be taken down, and for implementing that decision as appropriate.

2.7 We know that content moderation systems, particularly those deployed across a very large userbase, cannot provide a guarantee that users will not encounter any illegal content. However, well-designed and resourced content moderation systems and processes can significantly reduce that risk and help to protect users.

## Structure of this chapter[5]

2.8 In the next section, we explain the general approach we have taken to the U2U content moderation measures. We begin by outlining the approach we proposed in our November

[3] Policy Department for Economic, Scientific and Quality of Life Policies, 2020. Online Platforms' Moderation of Illegal Content Online: Laws, Practices and Options for Reform. [accessed 24 November 2024].
[4] 2017. Content Moderation. Encyclopedia of Big Data. [accessed 24 November 2024].

[5] We have an equivalent chapter in which we set out our recommendations on moderation on search services (chapter 3 of this Volume: 'Search Moderation').

2023 Illegal Harms Consultation ('November 2023 Consultation'), and then explain the approach we have decided to take based on the stakeholder feedback we received on this.

2.9 The following sections then explain in detail the eight measures we have decided on, and set out how we have considered them against the analytical framework.

## Our approach in the November 2023 Consultation

2.10 In our November 2023 Consultation, we considered three potential approaches to drafting the content moderation measures:

- **Approach 1** – specify in detail how providers should configure their content moderation systems and processes.
- **Approach 2** – specify in detail the outcomes content moderation systems and processes should achieve – for example, by setting detailed key performance indicators (KPIs) – but leave the design to providers.
- **Approach 3** – require providers to operate a content moderation system and (where relevant) set out the factors to which they should have regard when designing their content moderation systems and processes.

2.11 We proposed to pursue **Approach 3** because we considered this would allow providers greater flexibility to comply with the measures in ways that may be more proportionate and cost effective for them, while still setting out the important factors that providers should take into account where relevant. We considered that this approach was particularly beneficial given the diverse range of services in scope of the Act and the fast-moving pace of technological development.

2.12 We identified some areas in which to be more prescriptive and we set these out in chapter 4 of this Volume: 'Automated content moderation ('ACM')'. In that chapter, we recommend the use of ACM technology with a view to identifying further illegal content or suspected illegal content.

2.13 We proposed not to pursue **Approach 1** or **Approach 2** because:

- we did not have enough evidence to specify in detail every aspect of how providers should configure their content moderation systems and processes, or the outcomes that those systems and processes should achieve;
- there is no consensus on the optimum approach to content moderation;
- different approaches may be more appropriate in different circumstances and for different types of service; and
- taking a prescriptive approach at this stage would give rise to a substantial risk of regulatory failure and unforeseen consequences, which could lead to significant disruption in the sector – we considered that this could lead to potentially increased, rather than decreased, harm to users.

## Feedback on our approach

2.14 Several stakeholders expressed their overall support for our proposed content moderation measures.[6] Some noted that the measures were reflective of industry practice.[7]

2.15 Spotify and the Center for Data Innovation expressed their support for taking Approach 3 for our content moderation measures.[8] Several stakeholders, while not directly referencing Approach 3, expressed their support for the flexibility of the Codes.[9]

2.16 Logically argued that we should have taken a more prescriptive approach to the content moderation Codes, more aligned to Approaches 1 or 2 that we consulted on.[10] [✂].[11] Global Network Initiative argued that some providers would struggle with the absence of appropriate legal benchmarks on content moderation on which their compliance with our measures would be assessed.[12] In response to the May 2024 Consultation, the National Crime Agency (NCA) argued that we should set clearer minimum standards around proportionate investment in content moderation.[13]

2.17 Mid Size Platform Group and techUK argued that the proposed content moderation measures were too prescriptive.[14] Several stakeholders highlighted the particular challenges smaller service providers would face in complying with the measures.[15] Mid Size Platform Group highlighted that smaller service providers would face operational burdens

[6] Are, C. response to the November 2023 Illegal Harms Consultation, p.7; Children's Commissioner for England response to the November 2023 Illegal Harms Consultation, p.21; Dwyer, D. response to the November 2023 Illegal Harms Consultation, p.7; Evri response to the November 2023 Illegal Harms Consultation, p.5; LinkedIn response to the November 2023 Illegal Harms Consultation, p.9; Match Group response to the November 2023 Illegal Harms Consultation, p.9; Metropolitan Police Service and Counter Terrorism Policing response to the November 2023 Illegal Harms Consultation, p.2; Microsoft response to the November 2023 Illegal Harms Consultation, p.10; National Trading Standards eCrime Team response to the November 2023 Illegal Harms Consultation, p.8; Safecast response to the November 2023 Illegal Harms Consultation, p.7; South East Fermanagh Foundation response to the November 2023 Illegal Harms Consultation, p.9; Ukie response to the November 2023 Illegal Harms Consultation, p.15; WeProtect Global Alliance response to the November 2023 Illegal Harms Consultation, p.12; Welsh Government response to the November 2023 Illegal Harms Consultation, p.3.

[7] [✂]; Segregated Payments Ltd response to the November 2023 Illegal Harms Consultation, p.7.

[8] Center for Data Innovation response to the November 2023 Illegal Harms Consultation, p.10; Spotify response to the November 2023 Illegal Harms Consultation, p.14.

[9] ACT: the App Association response to the November 2023 Illegal Harms Consultation, p.10; Booking.com response to the November 2023 Illegal Harms Consultation, p.8; Name withheld 5 response to the November 2023 Illegal Harms Consultation, p.9; Federation of Small Businesses response to the November 2023 Illegal Harms Consultation, p.3. Global Partners Digital response to November 2023 Illegal Harms Consultation, p.13; Snap response to November 2023 Illegal Harms Consultation, p.9. We note that Federation of Small Businesses made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.6

[10] Logically response to November 2023 Illegal Harms Consultation, p.17.

[11] [✂].

[12] Global Network Initiative response to the November 2023 Illegal Harms Consultation, p.8. We note that Global Network Initiative made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.13.

[13] NCA response to May 2024 Consultation on Protecting Children from Harms Online, p. 10

[14] Mid Size Platform Group response to the November 2023 Illegal Harms Consultation, p.8, techUK response to the November 2023 Illegal Harms Consultation, p.5.

[15] BILETA response to the November 2023 Illegal Harms Consultation, p.19; Federation of Small Businesses response to the November 2023 Consultation, p.3; Global Network Initiative response to the November 2023 Illegal Harms Consultation, p.8; INVIVIA response to the November 2023 Illegal Harms Consultation, p.12; Mid Size Platform Group response to the November 2023 Consultation, p.8; Online Dating and Discovery Association response to the November 2023 Illegal Harms Consultation, p.2.

in complying with the measures, where large services that have the financial resources, operational capacity and robust practices to more easily navigate regulatory challenges, increasing the barriers to market entry.[16]

2.18 In response to the May 2024 Consultation on Protecting Children from Harms Online ('May 2024 Consultation'), several stakeholders suggested that we should take a more 'outcomes-based' approach to content moderation, naming the outcomes providers should achieve through content moderation, rather than specific content moderation practices providers should adopt.[17]

## Decision on general approach to content moderation

2.19 We have decided not to change our approach in response to feedback from some stakeholders that the measures are too prescriptive.[18] We consider that the measures allow providers to follow them in ways they find most appropriate and cost-effective. We believe this to be a proportionate approach that is appropriate for the diverse range of services within scope of the regulations.

2.20 We also consider that the reasoning provided at consultation for not taking a more prescriptive approach for these measures, or an entirely outcomes-based approach for these measures still stand. The reasons for which are set out above in paragraph 2.13. We consider that these reasons also explain why we have not chosen to set legal benchmarks for compliance with our measures, or defined minimum standards for investment in content moderation.[19] [20]

2.21 We have therefore decided to broadly adopt Approach 3 when designing our content moderation measures. However, consistent with the hybrid approach to designing our measures[21], we have provided more specificity where we consider it to be appropriate to do so.

# Measures on reviewing, assessing and swiftly taking down content

2.22 In the November 2023 Consultation, we proposed that providers should have systems and processes designed to swiftly take down content of which they are aware. For this purpose, we proposed that when providers have reason to suspect content is illegal content, they should either:

- make an illegal content judgement in relation to the content and, if they determine that the content is illegal content, swiftly take the content down;

[16] Mid Size Platform Group response to the November 2023 Consultation, p.8.

[17] Mid Size Platform Group response to the May 2024 Consultation on Protecting Children from Harms Online, p.9; Molly Rose Foundation response to May 2024 Consultation on Protecting Children from Harms Online, pp.40-41.

[18] Mid Size Platform Group response to the November 2023 Consultation, p.8; techUK response to the November 2023 Consultation, p.5.

[19] Global Network Initiative response to the November 2023 Consultation, p.8- we note that Global Network Initiative made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.13.

[20] NCA response to May 2024 Consultation on Protecting Children from Harms Online, p. 10

[21] See 'Our approach to developing Codes measures'.

- or where the providers are satisfied that their terms of service prohibit the types of illegal content which they have reason to suspect exist, consider whether the content is in breach of those terms of service and, if it is, swiftly take the content down.

2.23 The proposed measure was designed to reflect the safety duties in the Act.[22] In effect, these require that service providers must have proportionate systems and processes in place to moderate U2U content that is illegal content.

2.24 We proposed this measure should apply to all U2U service providers.

## Summary of stakeholder feedback[23]

2.25 In addition to those stakeholders who expressed broader support for the full package of content moderation measures outlined in paragraph 2.14, some also expressed support specifically for this measure.[24] [✂].[25] [✂].[26]

2.26 There were several areas where stakeholders felt the measure should be amended. These included (but were not limited to):

- the relationship between content moderation, terms of service, and the Illegal Content Judgements Guidance (ICJG);
- the part of the content moderation process in which providers should act 'swiftly';
- the measure's impact on freedom of expression rights; and
- the measure's impact on privacy rights.

2.27 We outline these stakeholder concerns in more detail in the following sections, and address additional stakeholder responses in Annex 1.

### The relationship between content moderation, terms of service and the Illegal Content Judgements Guidance (ICJG)

2.28 We received opposing stakeholder views on our proposal that providers can use either the ICJG or their terms of service to review content and make decisions about whether to take down content.

2.29 Meta expressed its support for the two options we suggested. It explained that it proposes to adopt an approach of assessing content against its respective terms of service, and then (where necessary) reviewing UK content for local illegality on a case-by-case basis.[27]

2.30 Microsoft and Snap expressed a preference for the option of a provider taking action against potentially illegal content by applying its own terms of service to that content.[28]

[22] The duties under Section 10 of the Act

[23] Note this list in not exhaustive, and further responses can be found in Annex 1.

[24] Federation of Small Businesses response to the November 2023 Consultation, p. 3; Match Group response to November 2023 Consultation, p.9-10; Molly Rose Foundation response to November 2023 Illegal Harms Consultation, p.35. We note that Meta (p.19), Children's Commissioner for England (p.59) and NICCY (p.32) expressed support for the equivalent measure to this in response to the May 2024 Consultation on Protecting Children from Harms Online.

[25] [✂].

[26] [✂].

[27] Meta response to November 2023 Illegal Harms Consultation, p.35.

[28] Microsoft response to November 2023 Consultation, p.20; Snap response to the November 2023 Illegal Harms Consultation, p.9.

2.31 Several stakeholders raised concerns that the option to assess content through terms of service would require providers to moderate content for all global users based on UK law.[29] In raising this concern, Snap requested some clarification on the level of granularity required in terms of service, suggesting that its terms of service apply globally and it would therefore be inappropriate to cover each illegal harm in UK legislation in granular detail.[30]

2.32 We address these points in paragraphs 2.47-2.51 in the section entitled 'How these measures work'.

2.33 Several stakeholders raised rights-related concerns about giving the option to providers to make illegal content judgements to assess whether they need to take content down. Snap suggested that to do this, providers would necessarily be speculative and would err on the side of over-reporting, chilling freedom of speech and infringing on users' privacy rights.[31] Big Brother Watch and Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) suggested the process of assessing content for illegality should be given to law enforcement and judicial authorities, instead of being outsourced to service providers, due to concerns about the ability of service providers to make illegal content judgements, and the risks this could pose for over-removal and freedom of expression.[32] The New Zealand Classification Office suggested that we should support providers to identify content and apply legal tests to "edge cases".[33]

2.34 We address these points in paragraph 2.80 under the section entitled 'Rights impact'.

### The part of the content moderation process where providers should act 'swiftly'

2.35 Some stakeholders commented on the fact that the obligation for providers to be "swift" applies at the stage of the moderation process after which providers are "aware" of content.

2.36 In response to a corresponding measure proposed in the May 2024 Consultation, the National Society for the Prevention of Cruelty to Children (NSPCC) raised concerns that the focus of the measure was solely on how providers should respond to content once they become aware of it (rather than introducing proactive or preventative measures to ensure providers are able to swiftly detect content).[34]

2.37 In contrast, Big Brother Watch supported our proposal that content moderation requirements should only apply to illegal content of which providers are aware. It raised concerns about freedom of expression if providers were required to 'prevent' illegal content through content moderation.[35]

2.38 In response to the November 2023 Consultation, several stakeholders misinterpreted which part of the content moderation process we were referring to when we recommend

[29]Meta response to November 2023 Consultation, p.35; Name Withheld 3 response to the November 2023 Illegal Harms Consultation, p.9; Snap response to the November 2023 Illegal Harms Consultation, p.9; Wikimedia Foundation response to the November 2023 Illegal Harms Consultation, p.28

[30] Snap response to the November 2023 Consultation, p.9.

[31] Snap response to November 2023 Consultation, p.9. We note Snap made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.24

[32] Big Brother Watch response to the November 2023 Illegal Harms Consultation, p.2; CELE response to the November 2023 Illegal Harms Consultation, pp.7-8.

[33] New Zealand Classification Office response to November 2023 Illegal Harms Consultation, p.7.

[34] NSPCC response to the May 2024 Consultation on Protecting Children from Harms Online, p.50.

[35] Big Brother Watch response to the November 2023 Consultation, p.2.

providers should act swiftly. In expressing its agreement with the measure, Action for Primates referenced evidence about the current lack of urgency with which users' reports on illegal harms are processed by providers.[36] Several stakeholders argued that our definition of 'swiftly' within the measure was vague, or that we should set timelines within which content should be removed.[37] The British and Irish Law, Education, and Technology Association (BILETA) suggested this ambiguity was likely to lead to legal disputes between users and providers, thus possibly increasing the workload of the courts.[38] In contrast, Snap and Airbnb supported our decision not to define 'swiftly' within the measure, arguing that the meaning of 'swift' in content moderation is context-dependent.[39]

2.39 We address these points in paragraphs 2.59 to 2.66 in the section entitled 'How these measures work'.

## Taking down content

2.40 A small number of services said that it is not currently technically feasible for them to take content down. For example, WhatsApp said that it is not able to delete message content stored on-device, or hosted on a third- party back-up server, including following a user report. It said that if content reported by a user to WhatsApp is determined to be policy-violating, WhatsApp will take appropriate action which may include banning individual group members or admins, disbanding a group or banning all members of a group.[40]

## Freedom of expression

2.41 Several stakeholders expressed concerns about the overarching impact the content moderation Codes would have on users' right to freedom of expression. Open Rights Group requested more information about how we plan to encourage providers to protect freedom of expression in the content moderation measures.[41] Global Partners Digital highlighted the disproportionate impact that content moderation can have on the freedom of speech of vulnerable and marginalised communities and urged us to conduct further research on the differential impact of the Codes on such communities and on ways to mitigate that risk. It also suggested that we should provide further incentives for accurate illegal content removal.[42] BILETA expressed concerns that measures to safeguard rights (including their freedom of expression, privacy and data protection rights) are insufficient, and do not protect user content from the removal of legitimate content which then needs to be addressed by the user.[43]

[36] Action for Primates response to the November 2023 Illegal Harms Consultation, p.7.
[37]BILETA response to the November 2023 Consultation, p.20; [✂]; EVAW Coalition response to the November 2023 Illegal Harms Consultation p.3; Refuge response to the November 2023 Illegal Harms Consultation, p.12. We note that Kooth (p.7) and NSPCC (p.50) made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.7.
[38] BILETA response to the November 2023 Consultation, p.20.
[39] Airbnb response to November 2023 Illegal Harms Consultation, p.14; Snap response to the November 2023 Consultation, p. 9. We note that Meta response to the May 2024 Consultation on Protecting Children from Harms Online p.20.
[40] Letter from WhatsApp dated 22 November. We note WhatsApp have also released similar information publicly. WhatsApp, no date. About reporting and blocking someone on WhatsApp. [accessed 26 November 2024].
[41] Open Rights Group response to November 2023 Illegal Harms Consultation, p.2.
[42] Global Partners Digital response to November 2023 Consultation, p.14.
[43] BILETA response to November 2023 Consultation, p.8.

2.42 We address these points in paragraphs 2.76 to 2.88 in the section entitled 'Rights impact'.

### Privacy

2.43 The Information Commissioner's Office (ICO) disagreed with the view set out in our impact assessment that this measure's impact on users' right to privacy would be slight.[44] It stated that the moderation of content using automated means will still have data protection implications for service users whose content is being scanned. It also argued that the privacy impact assessments we conducted for all the measures did not take sufficient account of the impact the proposed content moderation measures would have on information rights.[45]

2.44 We address these points in paragraphs 2.98 to 2.102 under the section entitled 'Rights impact'.

## Our decision

2.45 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation albeit with small amendments to reflect aspects of the feedback we received:

- We have divided the measure that we originally proposed into two measures:
  - > In the first measure, we recommend that providers should, as part of their content moderation function, have systems and processes designed to review and assess content the provider has reason to suspect may be illegal content. For this purpose, 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 in relation to the content or, if its terms of service prohibit the suspected types of illegal content, it should consider whether the content is in breach of those terms of service.
  - > In the second measure, we recommend that providers should, as part of their content moderation function, have systems and processes designed to swiftly take down illegal content, and/or illegal content proxy[46] of which they are aware unless it is currently not technically feasible for them to achieve this outcome.[47]
- We have made this amendment to account for feedback we received suggesting that some services are configured in such a way that it is not currently technically feasible for them to take down content.
- We have included additional references to privacy safeguards in these measures to clarify how privacy rights are protected by our measures.

2.46 The full text of the measures can be found in the Illegal Content Codes of Practice and they are referred to as measures ICU C1 and ICU C2. These measures will be included in our Codes of Practice for U2U services on terrorism, CSEA and other duties.

[44] ICO response to the November 2023 Illegal Harms Consultation, p.12.

[45] ICO response to the November 2023 Consultation, p.11.

[46] In the Codes, we define "illegal content proxy" as content that a provider determines to be in breach of its terms of service, where: the provider had 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.

[47] We have redefined 'content moderation function' to reflect that this encompasses both these measures. We now define this as the systems and processes designed to review, assess and take action in relation to content, including content a provider has reason to suspect may be illegal content.

# Our reasoning

## How these measures work

### The relationship between terms of service and illegal content judgements in content moderation

2.47 In the measures, we recommend that, 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 in relation to the content; or
- where the provider is satisfied its terms of service prohibit the types of illegal content which it has reason to suspect exist, consider whether the content is in breach of those terms of service.

2.48 We also consider it may be appropriate for a content moderation system to adopt an approach that combines the two processes above. For example, a provider could act in accordance with this measure by largely assessing content based on its terms of service to the extent that the terms of service prohibit the suspected types of illegal content, but make illegal content judgements in relation to the content on a case-by-case basis, where the terms of service do not capture all illegal content and there is reason to suspect (as set out in paragraph 2.53) that the content is illegal content.

2.49 We have given providers these options because we recognise that many service providers design their terms of service and community guidelines both to comply with existing laws in multiple jurisdictions and to meet their own commercial needs. For example, if a provider has already decided that it wishes to remove all bullying and harassment from its service, we do not consider that complying with the takedown duty creates a need for that provider to go on to make a potentially more complex judgement about whether this was racially or religiously aggravated or amounts to the criminal offence of harassment. Instead, the provider could simply apply its terms of service.

2.50 We therefore consider that service providers should have a choice. They may set about making illegal content judgements in relation to individual pieces of content for the express purpose of complying with the safety duties. In practice this would necessarily give effect to terms of service the provider adopts under section 10(5) of the Act (which sets out how users are to be protected from illegal content). The alternative is that they moderate illegal content by reference to provisions in their terms of service which would be cast broadly enough to necessarily cover illegal content.

2.51 Where a provider assesses content that it suspects to be illegal content against its terms of service (rather than making an illegal content judgement), this content would be an "illegal content proxy".[48]

2.52 We do not agree with stakeholders' feedback that the measure would require providers to apply UK rules on content globally.[49] The Act does not prevent service providers from having different terms of service for UK users and for users elsewhere in the world. In

[48] In the Codes, we define "illegal content proxy" as content that a provider determines to be in breach of its terms of service, where: the provider had 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.

[49] Meta response to November 2023 Consultation, p.35; Name Withheld 3 response to the November 2023 Illegal Harms Consultation, p.9; Snap response to the November 2023 Consultation, p.9; Wikimedia Foundation response to the November 2023 Consultation, p.28.

practice, where the Act requires content to be taken down, this refers to taking it down for UK users. Providers have flexibility over how their terms of service are drafted to cover illegal content, so long as they are drafted to specify how UK users are to be protected from illegal content.

2.53 We consider that providers may be alerted to content they suspect may be illegal content (as the Act defines it) in a variety of ways. The Act governs its treatment of complaints by UK users and affected persons, which we consider further in chapter 6 of this Volume: 'Reporting and complaints'. A complaint by a UK user or affected person about suspected illegal content is grounds to suspect the content may be illegal, except where the provider determines it to be manifestly unfounded as set out in chapter 6 of this Volume: 'Reporting and complaints'. In the same chapter, we also recommend a means for entities with appropriate expertise and information ('trusted flaggers') to report suspected illegal content to service providers. A report from a trusted flagger about matters within its expertise would always be such a reason. Additionally, in chapter 4 of this Volume: 'Automated content moderation', we discuss the ACM technology that we recommend providers use to identify further illegal content or suspected illegal content. Providers may choose to use other kinds of technology or human content moderators to identify content suspected to be illegal content as defined in the Act.

2.54 In the light of the responses to our consultation, we have considered carefully how this measure should apply to end-to-end encrypted services. End-to-end encrypted services are not able to proactively review content in the same way that other services are. However, a number of providers said that they can deal with complaints when the content is revealed to them by the user. WhatsApp emphasised that personal messages and calls between users on WhatsApp are end-to-end encrypted by default and that end-to-end encryption ensures only the user and the person they're communicating with can read or listen to what is sent, and nobody in between, not even WhatsApp. However, it also noted that when a user reports content to it, it receives up to the last five messages sent to the user by the reported sender or group. It also receives the reported group or user ID; information about when the message was sent, and the type of message sent such as an image, video or text.[50] This suggests that it is possible for providers of end-to-encrypted services to configure their services such that they can view and assess content which has been reported to them as being illegal.

2.55 In light of this, our measure recommends that all providers of user to user services, including end-to-end encrypted services, should, as part of their content moderation function, have systems and processes in place to review and assess content which the provider has reason to suspect may be illegal content (which in practice may only be possible where the user concerned reports it), and either make an illegal content judgement or a judgement about whether the content is in breach of their terms of service.

2.56 We consider applying the measure to end-to-end encrypted services in this way is of fundamental importance to the operation of the regulatory regime. Without access to at least a copy of content to which a user report refers, we consider that in the vast majority of cases it would be impossible for a provider to have reasonable grounds on which to make a judgement as to whether content is illegal, or otherwise to take a view on whether the

[50] Letter from WhatsApp dated 22 November. We note WhatsApp have also released similar information publicly. WhatsApp, no date. How to block and report someone. [accessed 26 November 2024].

content is violative of its terms of service.[51] The overall purpose of the Act is to make the use of regulated internet services “safer for individuals in the United Kingdom”.[52] Generally speaking, it would not make users safer if providers were to operate their services in a way which meant that almost every complaint they received about illegal content necessarily led to a decision that too little information was available to determine that the content was violative.

2.57 Therefore, to benefit from the safe harbour of the Codes, providers of end-to-end encrypted services should operate their complaints processes in a way that ensures they can have regard to reasonably available relevant information,[53] including the content complained of when making judgements about content. We are not recommending that providers should break end-to-end encryption to do this. However, as some providers of end-to-end encrypted services told us that they are able to make content judgements in a privacy preserving manner, we consider it is proportionate to recommend these services make this type of judgement.

### Swiftly taking down content of which providers are “aware”

2.58 We recommend that once providers have carried out the steps to review content described in paragraph 2.47, if the content is violative, most providers should take that content down swiftly.[54]

2.59 We consider that the responses which interpreted our proposed recommendation to be that providers should act swiftly in processing user reports may reflect a misunderstanding of our proposal: the obligation for a provider to act swiftly starts after a provider has reviewed content.

2.60 In the Act, there is a separate duty to report child sexual exploitation and abuse (CSEA) content to the National Crime Agency, which also relates to CSEA content of which the provider is “aware”. It is therefore important to be clear at what point, following a complaint or other alert, a provider should be considered to be “aware” that content is (rather than may be suspected to be) illegal content.

2.61 Due to the privacy implications of such reports, together with the potential impact of incorrect complaints on scarce public resources, we consider the reporting duty must arise at the point at which a provider has made a decision on the content, not immediately on receipt of the complaint. It follows that in section 10(3)(b) of the Act, the duty to take content down swiftly must refer to the period in between the decision being made that the content is violative and the content being taken down.

2.62 We do not consider it necessary to provide detailed guidance, as recommended by some stakeholders, on how quickly a provider has to remove content once aware of it, in order to comply with this measure.[55] We expect providers will have systems and processes in place

[51] It would require enough complaints to accumulate from enough different and unrelated people for it to be reasonable to infer that a new complaint was neither malicious nor erroneous. To draw such an inference would require a substantial amount of harm to have already occurred.

[52] Section 1 of the Act.

[53] As required in Section 192 of the Act.

[54] We note that relevant moderation actions will be different for providers for whom it is currently not technically feasible to take down content.

[55]BILETA response to November 2023 Consultation, p.20; [✂]; EVAW Coalition response to November 2023 Consultation p.2; Refuge response to the November 2023 Consultation, p.12. We note that Kooth (p.7) and

to take down content they have assessed to be illegal as quickly as is feasible for their service and the content involved.

2.63 We disagree with BILETA's argument that our decision not to define "swiftly" would lead to disputes between users and providers and increase the workloads of the courts.[56] Users do not have a private right of action under the Act to bring a claim against a provider for breach of their safety duties.

2.64 More generally, and as set out in our explanation of why we have taken the approach we have to our measures as a whole, we consider that recommending time limits for the processing of reports of all suspected illegal content at this stage, as suggested by some stakeholders, could pose significant risks to user safety.[57] We provide further reasoning for not being more prescriptive on this in paragraphs 2.205, when explaining why we are not being more prescriptive in our measure on performance targets. We also accept that providers may better protect users from illegal content by prioritising certain types of suspected illegal content for review. This includes content that is likely to be illegal, content which is likely to be seen by a lot of people and/or content which, if illegal, would be particularly harmful. Applying an appropriate prioritisation process to content moderation, as we recommend in the measure on a policy for the prioritisation of content for review, may mean that some items of suspected illegal content are not reviewed as promptly as others. However, we consider that users may be better protected as a result of such prioritisation.

2.65 We consider that the safety duties about illegal content which require providers to have proportionate systems and processes to swiftly take down content of which they are aware mean that providers should also have in place systems and processes to consider content when alerted to its presence.[58] We have recommended other Codes measures which (while not setting time limits for the review of content) we consider will secure that providers take appropriately quick action before they are "aware" of content (as appropriate for their service). In chapter 6 of this Volume: 'Reporting and Complaints', we recommend measures relating to the handling of complaints. We also recommend that providers give complainants an indication of how long it will take for their complaint to be considered. In this chapter, we recommend measures on performance targets and resourcing.

2.66 We are not, at this stage, recommending all providers proactively detect all types of illegal content swiftly, before they have reason to suspect content may be illegal content as suggested by the NSPCC.[59] We recommend technology some providers should use to proactively detect content related to certain types of illegal harm in chapter 4 of this Volume: 'Automated content moderation'. However, we have not yet fully assessed what actions, if any, it would be proportionate to recommend providers of U2U services should take to detect other types of illegal content prior to having reason to suspect it is illegal.[60] It would therefore not be proportionate, at this stage, to set expectations for how quickly providers should detect content on their service using proactive technology. We recognise

NSPCC (p.50) made a similar point in response to the May 2024 Consultation on Protecting Children from harms online.

[56] BILETA response to November 2023 Consultation, p.20.

[57] Refuge response to November 2023 Illegal Harms Consultation, p.12.

[58] Section 10 of the Act.

[59] NSPCC response to the May 2024 Consultation on Protecting Children from Harms Online, p.50.

[60] For example, we recommend that certain providers use hash matching technology to detect CSAM.

that many providers proactively take steps to detect harmful or illegal content, and we welcome this. We are currently considering evidence surrounding the use of automated tools to proactively detect illegal content and the content most harmful to children, going beyond the automated detection measures we have already consulted on. We intend on consulting on these additional measures in Spring 2025.

## Benefits and effectiveness

2.67 We consider that there are clear benefits of these measures for protecting users from illegal content. Having effectively enforced content moderation systems and processes is one of the most important ways in which service providers can reduce the risk of users encountering illegal content.[61] Conversely, a lack of effective and consistently applied content moderation processes can lead to an increased risk of illegal content being present on services and subsequent harm to users.[62]

2.68 We consider that providers swiftly taking down illegal content of which they are aware has important benefits for user safety. If providers are aware that content is illegal, they should remove it swiftly to minimise further harm to users caused by the content or its amplification.

## Costs and risks

2.69 The costs of implementing these measures will vary from service to service. For providers of smaller low-risk services that receive few complaints, the costs could be low. Such services will only have a limited amount of content to review and are unlikely to require a complex content moderation system to do so effectively. They may have a process to assess all complaints regarding potentially illegal content as they arise and take down any illegal content to meet the minimum requirement of the Act. This may entail some small one-off costs of designing and implementing such a system. Ongoing costs associated with moderators reviewing the content and actioning where appropriate are likely to vary in proportion to the size and risk level of a service and are therefore expected to be small for small low-risk services.

[61] For example, there is evidence of increased user safety and a reduction in illegal (or harmful) content when investment is put into improving content moderation systems. Reddit, 2022. 2022 Transparency Report. [accessed 24 November 2024]

[62] For example, a report by the Institute for Strategic Dialogue (ISD) suggests that 'extreme right-wing activists' may view services with less moderation as preferable spaces for extremist discussions which may include illegal terrorist and hate content, when compared to services with more moderation. The Institute for Strategic Dialogue, 2021. Gaming and Extremism: The Extreme Right on Twitch. [accessed 24 November 2024]. A report by HOPE not hate and the Antisemitism Policy Trust suggested that minimal moderation on one messaging app (along with its "commitment to secrecy… and relative ease-of-use") has "lowered the hurdle for engaging in the politics of hate and has enabled extremist networks to propagandise, network and organise", saying the service could be "a powerful radicalisation tool, as individuals can quickly become immersed in bubbles practically free from moderation in which they receive constant streams of propaganda." HOPE not hate and the Antisemitism Policy Trust, 2021. Antisemitism and Misogyny: Overlap and Interplay. [accessed 24 November 2024]. There is also evidence of content moderation systems failing to tackle illegal harms or being used to facilitate illegal offences. House of Commons Home Affairs Committee, 2017, Hate crime: abuse, hate and extremism online. [accessed 24 November 2024]; Counter Extremism Project, 2018. OK Google, Show Me Extremism: Analysis of YouTube's Extremist Video Takedown Policy and Counter-Narrative Program. [accessed 24 November 2024] Amnesty International, 2022. Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya. [accessed 24 November 2024]. Ofcom, 2022. The Buffalo attack: Implications for online safety. [Accessed 24 November 2024]

2.70 Providers of larger and riskier services will typically face higher costs to develop content moderation systems and processes in line with these measures. These costs are likely to include both one-off costs of developing a system and ongoing costs of maintaining it. One-off costs for providers that decide to build their own systems internally may include hiring experienced content moderation systems designers, developing content moderation tools, project management and integration with data analytics/measurement software. For providers which are not building their systems internally, the main cost would be the adoption of third-party moderation solution and integration with their internal policies, tools, and processes, as well as the fees that they pay the third party to moderate for them. There would also be several ongoing costs relating to systems maintenance, hosting and data logging which would vary by service as well as the ongoing costs relating to detecting, reviewing, and taking down illegal content.

2.71 We consider that the costs discussed here reflect the base level of cost which is required to design and operate a content moderation system to review, assess and take down illegal content. We consider that a proportionate approach for large and risker services will also entail costs additional to this, as set out in the additional measures below.

2.72 Overall, we expect that the costs of implementing these measures will vary widely between services. For providers of the smallest low-risk services, costs are likely to be negligible or in the small thousands at most. For providers of some large or risky services these costs could extend to multiple millions depending on the approach taken, the volume of content on the service and/or the volume of reports received.

2.73 While the costs described in this section may be significant for some providers, we consider that these measures capture the minimum steps to ensure that they meet their duty to have proportionate systems and processes designed to minimise the length of time for which any illegal content is present and to take down illegal content of which they become aware having been alerted to it by another person or in any other way. They are also the minimum needed to enable providers to consider complaints about illegal content appropriately. Incurring these costs is therefore necessary to meet the requirements of the Act.

2.74 Providers of some end-to-end encrypted services will have to make changes to the design of their complaints processes in order for them to consider complaints effectively. As set out above, we understand that it is possible for such providers to design their complaints processes to do this. This will entail one-off system design costs, and ongoing maintenance costs. It would not require service providers to break end-to-end encryption. We are recommending that service providers ensure that their complaints procedures enable complainants to share potential illegal content with them, which we do not consider to be a fundamental change to the nature of these services. The fact that some end-to-end encrypted services can already consider complaints effectively suggests that the costs are not prohibitive. In any event, we consider that being able to consider complaints in this way is so fundamental to the effective operation of the regulatory regime that the costs are proportionate.

## Rights impact

2.75 In implementing these measures, service providers may take content moderation steps which have a potentially significant impact on the rights of users, in particular, their rights under the European Convention on Human Rights (ECHR) to privacy (Article 8), freedom of expression (Article 10) and freedom of association (Article 11).

### Freedom of expression and freedom of association

2.76 As explained in 'Introduction, our duties and navigating the Statement', Article 10 of the ECHR sets out the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 11 sets out the right to associate with others. We must exercise our duties under the Act in light of users' and services' Article 10 and 11 rights and not interfere with these rights unless we are satisfied that to do so is prescribed by law, pursues a legitimate aim, is proportionate to the legitimate aim, and corresponds to a pressing social need.

2.77 With these two measures, potential interference with users' freedom of expression arises where the service provider decides to apply content moderation processes to material it suspects may be illegal content, as in this case the service provider may need to restrict users' access to it. This impact is potentially significant if that judgement is incorrect (as in this case, there would not be a substantial public interest in access to the piece of content in question being restricted).

2.78 These measures involve service providers reviewing people's communications, which may also have an impact on their freedom of association. We also note that some service providers choose to restrict or remove users' ability to use their services, if they are found to have shared illegal content; however, this is not a result of these measures.

2.79 The duty for services to treat illegal content appropriately is a requirement of the Act, and not of these measures. The measures do not involve services taking any action against illegal content of which they are not yet aware, or to restrict access to any content which they do not judge to be illegal content, or which is not in breach of their terms of service which the provider is satisfied prohibit the types of illegal content defined in the Act. The measures are not prescriptive about how such content is to be moderated, instead they seek to secure that the provider's systems and processes are designed so that they take steps to swiftly take down illegal content of which they are aware. To the extent that the actions taken as a result of these measure affect users' ability to access or share illegal content, we consider that is justified in line with the duties of the Act, as the benefits of the protections to users should outweigh the restrictions on other users' rights to encounter or share such content.

2.80 In relation to concerns from stakeholders about providers (rather than law enforcement or judicial authorities) having the option to make illegal content judgements, we note that this is the scheme of the Act. It is not practical for us to be consulted on content moderation decisions for every service within scope of the Act (which we predict to amount to over 100,000 services). Our role under the Act is to regulate systems and processes, rather than individual pieces of content.

2.81 There is a potential risk of error in content moderation, for example where a provider makes an incorrect judgement as to the illegality of content. Impacts on freedom of expression could in principle arise in relation to the most highly protected forms of speech, such as religious expression (which could also affect users' rights to religion or belief under Article 9) or political speech, and in relation to kinds of content that the Act seeks to protect, such as content of democratic importance, journalistic content, and content from

Recognised News Publishers.[63] We recognise that, in certain circumstances, it can be difficult to assess whether such kinds of content should be classified as illegal content, especially when considering whether such content may constitute a threat, abuse, harassment or hate offence. We note the concerns raised by Global Partners Digital in relation to the disproportionate effects on the freedom of expression of vulnerable and marginalised groups.[64] We also note the risk it identifies that the definition of illegal content under the Act will result in an increase in content to be moderated, leading to more errors.[65]

2.82 However, the definition of illegal content is statutory. Providers have incentives to limit the amount of content that is wrongly actioned, to meet their users' expectations and to avoid the costs of dealing with appeals. Our measures on appeals (from paragraph 6.302 of chapter 6 of this Volume: 'Reporting and complaints') therefore act as a safeguard for freedom of expression. We have prepared the ICJG with careful regard to rights of freedom of expression and encourage service providers to have regard to the ICJG when implementing this measure, to assist with correctly identifying when freedom of expression considerations are particularly relevant to availability of certain content.

2.83 While it is not a requirement of these measures, we note that a greater degree of interference with users' rights could arise if the service provider chose to adopt terms of service which defined the content in relation to which users' access should be restricted more widely than is necessary to comply with the Act. In this case, services could also be restricting users' access to certain types of content which is not required under the duties in the Act, and might also not be harmful, or might be less severely harmful, to them. However, it remains open to service providers as a commercial matter (and in the exercise of their own right to freedom of expression) to decide what forms of content to allow or not to allow on their service so long as they comply with the Act. We have no power to prevent them from doing so. If they choose to do so more as a result of the Act, this is the effect of the Act and not our recommendations. However, service providers have incentives to meet their users' expectations in this regard, too.

2.84 The use of content moderation to limit users' exposure to illegal content could also have significant positive impacts on the freedom of expression and freedom of association rights of users and affected persons. More effective moderation of illegal content could result in safer spaces online where users may feel more able to join online communities and receive and impart (legal) ideas and information with others.

2.85 We have considered if there could be a risk of a more general effect on freedom of expression if UK users were, as a result of these measures, to cease to use well-moderated services. However, we do not consider that any such effect would be likely to arise given that the measure relates to illegal content. Many UK users already use service providers which have content moderation processes.

2.86 These measures may also have an impact on service providers' rights to freedom of expression as, to the extent that they do not already prohibit illegal content, they would need to take steps to ensure it is appropriately dealt with. However, this arises from the duties in the Act, and we are allowing flexibility as to the precise approach providers take.

[63] See the duties set out in sections 17, 18 and 19 of the Act.

[64] Global Partners Digital response to November 2023 Consultation, p.14.

[65] Global Partners Digital response to November 2023 Consultation, p.14.

We therefore consider that to the extent that these measures impact on services' rights to freedom of expression, it is likely to constitute the minimum degree of interference required to secure that service providers fulfil their safety duties about illegal content under the Act.

2.87 These measures also specify other Codes measures as safeguards for users' freedom of expression, in particular other content moderation measures, enabling users to complain if their content has been taken down on the basis that it is illegal content, and the policies and processes for complaints. These other Codes measures help to safeguard users' freedom of expression in a number of different ways, including ensuring that (where those other measures apply to the service in question) the service provider sets internal content policies and provides training and material to individuals working in moderation, which would support them in determining whether detected content has been accurately identified, and in providing a level of transparency for users about any technology used and how to make a complaint. Additionally, in accordance with the principles of the Act[66] and our duties under the Human Rights Act 1998,[67] we will have regard to the importance of freedom of expression and association when making any decisions about enforcement in relation to this measure, which acts as an additional safeguard for these rights.

2.88 Overall, and taking the benefits to users and affected persons into consideration, we consider that any impact on rights of freedom of expression and association from these measures is proportionate.

### Privacy

2.89 As explained in 'Introduction, our duties, and navigating the Statement', Article 8 of the ECHR sets out the right to respect for individuals' private and family life. An 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.

2.90 All content moderation, whether by automated tools or human moderators, will impact on the rights of individuals to privacy and their rights under data protection law (discussed further from paragraph 2.89 below). The degree of interference with the right to privacy will depend to a degree on the extent to which the nature of their affected content and communications is public or private, or, in other words, gives rise to a legitimate expectation of privacy.

2.91 These measures are not limited only to content or communications that are communicated publicly, and may lead to the review of content or communications in relation to which individuals might reasonably expect privacy.[68] This would involve more significant privacy impacts than moderation of content and communications that are widely publicly available (whether on the service concerned or more generally). The impact on users' rights would

[66] In particular, see section 1(3)(b).

[67] Section 6.

[68] In the November 2023 Consultation, we consulted on draft guidance on content communicated 'publicly' and 'privately' under the Online Safety Act. Our final guidance recognises that whether content is communicated 'publicly' or 'privately' for the purposes of the Act will not necessarily align with whether that content engages users' (or other individuals') rights to privacy under Article 8 of the European Convention on Human Rights. For example, it is possible that users might have a right to privacy under Article 8 of the ECHR in relation to content which is communicated 'publicly' for the purposes of the Act. Conversely, users may not have a right to privacy under Article 8 of the ECHR in relation to content which is nevertheless communicated 'privately' for the purposes of the Act.

also be affected by the nature of the action taken as a result of the content moderation process. For example, the level of intrusion and significance of the impact is likely to be higher where content is judged to be violative. It is likely to be particularly high when content its judged to be CSEA content because this triggers the reporting duty under section 66 of the Act.

2.92 The privacy implications of the measure related to illegal content judgments apply with further force in the case of end-to-end encrypted services, since users of the services may expect that the contents of their communications will only be visible to the recipient. However, we consider that applying the measure in the way we have represents the minimum possible interference to secure user safety. The purpose of the Act is to make the use of regulated internet services (including end-to-end encrypted services) safer for users in the UK. It is in this context that the proportionality of this measure must be considered and in which a balance must be achieved between keeping individuals safe from illegal content and rights to privacy. Were providers of end-to-end encrypted services not making illegal content judgements, or if nearly all complaints received about illegal content could not be upheld due to a lack of information about the content, then in practice there would be little value in such a provider complying with its statutory duty to have a reporting function and a complaints procedure and users in the UK would not be made safer. We consider it would defeat the purpose of the Act if providers were considering complaints but could not determine that content was illegal content. The measure does not recommend the provider to allow access by any third party. We therefore consider that the impact of this measure on privacy rights, in relation to the recommendation for providers of end-to-end encrypted services to make illegal content judgements, is proportionate. Overall, and taking the benefits to users and affected persons into consideration, we consider that any impact on privacy rights from this measure is proportionate.

2.93 The removal of some kinds of illegal content also acts directly to protect the rights of victims, for example those depicted in child sexual abuse material (CSAM) or content that amounts to intimate image abuse. This sort of content causes ongoing harm to victims from knowing that the material continues to circulate online (or in some cases themselves viewing that material), or from being identified by persons who have viewed that material. Its removal protects victims' and survivors' rights under Article 8 ECHR and protects their personal data.[69]

2.94 Where CSAM is identified through the operation of the content moderation function recommended by this measure, providers may be required (or choose) to report this to a law enforcement authority or to a designated reporting body. Relevantly, section 66 of the Act (which is not yet in force) sets out duties for search service providers to report to the National Crime Agency (NCA) detected CSEA content which is not otherwise reported. Providers may also have additional CSEA reporting duties in other jurisdictions or have voluntary reporting arrangements. Aspects of the Act's reporting duties are to be further

[69] Review by human moderators of content accurately detected to be CSAM also represents a significant interference with the privacy rights of the victims it depicts. However, that review forms an important part of ensuring that these measures are proportionate and appropriate for service providers to take for the purposes of complying with their illegal content safety duties. We therefore consider that the intrusion into victims' privacy rights is necessary, and that no less intrusive approach would be a suitable alternative.

defined in regulations made by the Secretary of State.[70] However, a report may include information about identifiable individuals (for example, victims or perpetrators who appear in that content), which may present an additional risk to the right to privacy.

2.95 In part, any such interference results from the reporting duties created by the Act or by existing legislation in other jurisdictions. Where CSEA content identified by a service provider is correctly reported in line with the Act, any interference is prescribed by the relevant legislation. In enacting the legislation, Parliament has already made a judgement that such interference is a proportionate way of securing the relevant public interest objectives. However, we recognise that the risk to the privacy rights of individuals will be particularly acute in respect of any content that is incorrectly reported. In that regard, we consider that the accuracy principle in data protection law is of particular relevance in the context of reporting CSEA content (as reiterated in paragraph 2.99 below).

2.96 The duty for services to treat illegal content appropriately, including through the application of content moderation systems and processes, is a requirement of the Act, and not of these measures, and we are giving services flexibility as to precisely how they implement this and what action they take. We recognise that depending on how service providers decide to implement these measures, it could result in a greater or lesser impact on users' privacy rights. However, as noted above, it remains open to services in the exercise of their own rights to freedom of expression to decide what forms of content to allow or not to allow on their service, and what forms of personal data they consider they need to gather to enforce their content polices, so long as they comply with the Act and the requirements of data protection legislation.[71] Providers are also required by the Act to have particular regard to users' privacy rights when deciding on and implementing safety measures.[72]

2.97 Overall, and taking the benefits to users and affected persons into consideration, we consider that any impact on privacy rights from these measures is proportionate.

### Data protection

2.98 The degree of impact will also depend on the extent of personal data about individuals which may need to be processed. These measures do 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; we give guidance about that separately in our illegal content judgments guidance. We consider that service providers can implement these measures in a way which minimises the amount of personal data which may be processed or retained so that it is no more than needed to give effect to their moderation processes.

2.99 Providers should familiarise themselves with applicable data protection legislation and relevant guidance from the ICO to understand how to comply with the UK data protection regime in processing users' personal data for the purposes of this measure.[73] This means they should apply appropriate safeguards to protect the rights of users, including for

[70] Section 67 of the Act requires the Secretary of State to make regulations which will set out the information to be included in reports to the NCA and may also require the retention of user-generated content, user data and associated metadata.

[71] Ofcom has given guidance on what information we consider to be reasonably available to service providers for the purposes of making illegal content judgments, in the preparation of which we have had regard to the right to privacy and the principle of data minimisation.

[72] Set out in section 22 of the Act in relation to U2U services.

[73] Such as UK GDPR guidance and resources and Content moderation and data protection.

example having regard to the need for personal data to be accurate. We note that this may be particularly relevant in the context of the provider reporting any CSEA content identified to the NCA or other law enforcement agency as described in paragraphs 2.90 to 2.94 above. Providers may also use third parties to carry out content moderation on their behalf. ICO guidance is clear that where third parties are used, it is for the service provider and that third party to identify their respective roles and obligations under data protection law and ensure that all the requirements of data protection law are met.[74]

2.100 Insofar as providers use automated processing in content moderation (besides the measures we consider in chapter 4 of this Volume: 'Automated content moderation'), they should refer to ICO guidance on content moderation where applicable to determine whether the processing is solely automated i.e. has no meaningful human involvement, and results in decisions that have a legal or similarly significant effect on users.[75] We consider that the safeguards provided for under applicable data protection legislation and explained in the guidance from the ICO will help to ensure that the impact of any automated processing on data protection and privacy rights is minimised.

2.101 Further to feedback from the ICO, we have also updated these measures to include specific references to the privacy safeguards provided by other measures which apply to certain providers operating a content moderation function.[76] We consider this clarifies the protections afforded to individuals by the Codes and how this measure seeks to minimise the impact on individuals' privacy rights.

2.102 Overall, we consider that (assuming service providers also comply with applicable data protection legislation requirements and guidance) the impact of these measure as a result of services' content moderation decisions and processes on users' rights to privacy and data protection rights, above and beyond the requirements of the Act, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their safety duties about illegal content under the Act. Taking this, and the benefits to users and affected persons into consideration, we consider that any impact on data protection rights from these measures is proportionate.

## Who these measures apply to

### Measure on reviewing and assessing content

2.103 All providers are required by the Act to have a complaints handling process, and section 21 requires appropriate action to be taken by the provider in response to relevant kinds of complaints. Complaints are an important way in which service providers can learn about harms on their service (and for some providers may be the only way). Any recommendations we make in Codes about identifying suspected illegal content are recommendations we make because we consider that they are an appropriate way to comply with the safety duty.

2.104 We therefore consider that the minimum required by the Act is that providers be equipped to consider content that has been flagged to them by complaints under the Act, or which has come to their attention because of measures we have recommended. Identifying illegal

[74] Further information on the requirements for contracts between data controllers and processors can be found at Contracts and liabilities between controllers and processors. See also ICO Guidance on controllers/processors.

[75] In which case Article 22 UK GDPR requirements are likely to apply.

[76] ICO response to November 2023 Consultation, p.12.

content will benefit users as it allows providers to understand harms on their services and is a necessary step to taking appropriate action on such content, including taking it down.

2.105 We therefore consider that the measure on determining whether suspected content is illegal or in breach of the service's terms of service is proportionate for all U2U service providers.

### Measure on taking down content swiftly

2.106 The Act also requires a provider to have proportionate systems and processes designed to, where the provider is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way, swiftly take down such content. Section 10(3)(b) of the Act links complaints to a takedown duty in a way that suggests it is not open to providers to ignore complaints about content they judge to be illegal (where it is technically possible for them to take down the content they are aware is illegal). Taking down illegal content reduces users' exposure to it, preventing harm to them; it also reduces the ongoing harm which may otherwise be caused to those depicted in some kinds of illegal content. We therefore start from the position that all providers should have systems and processes to take down illegal content of which they are aware (where it is technically possible for them to do this).

2.107 Our approach is to recommend that providers have proportionate systems and processes designed to take down illegal content swiftly, but without specifying how this is done. We consider that the impact the measure has on services is mitigated by the flexibility of this measure, as we are not being prescriptive as to how providers implement content moderation systems and processes, allowing providers to take cost-effective processes that are proportionate to the context of each service. We expect that small services which are low-risk for all kinds of illegal harm can appropriately review and take down content using simpler, less costly systems and processes, and the moderation costs to a small service that receives very few or no user reports are expected to be minimal.

2.108 We have therefore decided to apply the measure on taking down illegal or violative content of which the service is aware swiftly, to all providers of U2U services, subject to the proviso explained below.[77]

2.109 As set out above, evidence presented by stakeholders and our own technical analysis shows that a relatively small minority of U2U services in scope of the Act are configured in such a way that it is currently technically infeasible for them to take down content. The Act states that Ofcom must have regard to the principle that the measures in our Codes must be proportionate and technically feasible. As such, although the measure under discussion applies to providers of all user to user services, the text of the recommendation itself makes clear that we do not expect services to take down content where it is not currently technically feasible for them to achieve this outcome.

2.110 Where a service provider claims that it is technically infeasible for it to take down content we will investigate this. Should we then find that it is technically feasible for the provider to take content down the measure will apply to them. We do not consider that technical limitations will necessarily remain on an ongoing basis. Given the importance of this

[77] This measure applies to all providers of U2U services. However, as explained below the text of the recommendation makes clear that we do not expect services to take down content where it is not currently technically feasible for them to achieve this outcome.

measure, we expect providers to invest in the development of new technologies to keep users safe whilst protecting user privacy.

2.111 Any provider that can currently take down content 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 transparently to explain to Ofcom the nature of the risks arising from this decision and plans for mitigating these risks.

## Conclusion

2.112 Our analysis shows that the measures we are recommending are likely to deliver significant protections for users from harm and that the costs and impact on rights that will result from them are proportionate. Therefore, we have decided to recommend the measure broadly as proposed in our November 2023 Consultation except for some small amendments. We have divided it into two different measures so as to distinguish between reviewing and making judgements about content and taking action following the judgment. We have added references to safeguards for privacy rights to these measures.

2.113 We therefore recommend that all U2U service providers should have systems and processes designed to review and assess content the provider has reason to suspect is illegal content. We also recommend that all U2U providers have systems and processes designed to swiftly take down illegal content of which they are aware on their services, unless it is currently not technically feasible for them to achieve this outcome.[78]

2.114 These measures will be included in our Illegal Content Codes of Practice for U2U services on terrorism, CSEA and other duties. They are referred to within these Codes as ICU C1 and ICU C2.

# Measure on internal content policies

2.115 In our November 2023 Consultation, we proposed that providers of large services and providers of multi-risk U2U services should set internal content moderation and have regard to the findings of their risk assessment and any evidence of emerging harms on the service when doing so.

2.116 In our proposed amendments to the Illegal Content Codes (which we consulted on in May 2024 alongside the Protection of Children measures), we altered the reference to "emerging harm" and instead recommended that providers should have processes in place to update these policies in response to "any evidence of new and increasing illegal harm on the service (as tracked in accordance with Measure ICU A5 in Volume 1: chapter 5: 'Governance and accountability').

2.117 We said that where services are large or multi-risk, it is important that providers have clear content moderation policies in order to ensure consistency, accuracy, and timeliness of decision-making, because they may need to moderate large volumes of diverse content.

[78] See paragraphs 2.106 to 2.111 above, where we set out our reasons for the approach.

## Summary of stakeholder feedback[79]

2.118 In addition to those stakeholders who expressed broader support for the full package of content moderation measures outlined in paragraph 2.14, some expressed further support specifically for this measure.[80]

2.119 There were several areas where stakeholders felt this measure should be amended. These included (but were not limited to):

- the publication of internal content policies;
- what should be included in internal content policies; and
- who the measure applies to.

2.120 We outline these stakeholder concerns in more detail in the following sub-sections, and address additional stakeholder responses in Annex 1.

### Publication of internal content policies

2.121 The Electronic Frontier Foundation said that it is imperative that internal standards are consistent with external standards made available to users.[81] In contrast, Snap agreed with our proposal that providers should not be obliged to publish internal content policies, due to the risk of training perpetrators on how to evade enforcement.[82]

2.122 We address these points in paragraph 2.133 under the section entitled 'How this measure works'.

### What should be included in internal content policies

2.123 Glitch suggested that we should be more specific in what should be included in internal content policies. It highlighted that the measure did not mention the importance of gender-sensitive moderation policies, and argued that without such policies, there is a risk of providers overlooking or downplaying gender-based harm in content moderation efforts.[83]

2.124 We address this point in paragraph 2.140 under the section entitled 'Benefits and effectiveness'.

### Who this measure applies to

2.125 Several stakeholders argued that this measure should apply to all providers of services with a specific risk, in addition to providers of large and multi-risk services.[84]

[79] Note this list in not exhaustive, and further responses can be found in Annex 1.

[80] Born Free Foundation response to November 2023 Illegal Harms Consultation, p.5; Cats Protection response to November 2023 Illegal Harms Consultation, p.10; Global Partners Digital response to November 2023 Illegal Harms Consultation, p.13; Meta response to November 2023 Consultation, p.22; [✂]; We note that Meta (made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.21.

[81] Electronic Frontier Foundation response to November 2023 Illegal Harms Consultation, p.9. We note that Big Brother Watch made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, pp. 37-38.

[82] Snap response to November 2023 Consultation, p.10.

[83] Glitch response to November 2023 Illegal Harms Consultation, p.6. We note that VAWG Sector Experts made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.12.

[84] Age Verification Providers Association response to November 2023 Illegal Harms Consultation, p.2; NSPCC response to November 2023 Illegal Harms Consultation, p.20; VerifyMy response to November 2023 Illegal Harms Consultation, p.6; Yoti response to November 2023 Illegal Harms Consultation, p.16.

2.126 Barnardo's disagreed with our recommendation that only some services should set internal content policies, and said that illegal content, including CSEA, can be present on any service, no matter the size or perceived risk.[85] The Online Safety Act Network (OSA Network) suggested that there was no evidence underpinning our assessment that services which are small and low-risk are unlikely to face large volumes of content they need to assess.[86] We note that in response to the May 2024 Consultation, some stakeholders said that an equivalent measure should apply to all service providers.[87]

2.127 We address these points in paragraph 2.156 to 2.157 under the section entitled 'Who this measure applies to'.

## Our decision

2.128 We have decided to broadly confirm the measure as proposed in the November 2023 Consultation, including the subsequent amendment consulted on alongside the May 2024 Consultation. We have made one minor clarificatory change:

- Our measure now says that in setting and recording internal content policies, providers should have processes in place for updating these policies in response to evidence of new and increasing illegal harm on the service (as tracked in accordance with Measure ICU A5 in Volume 1: chapter 5: 'Governance and accountability'). This is to clarify that we are not recommending providers update their internal content policies every time they receive evidence of new and increasing illegal harm on their services, but that they have processes in place to do so where appropriate.

2.129 The full text of the measure can be found in our Illegal Content Codes of Practice for U2U services and is referred to as ICU C3. This measure will be included in our Codes of Practice for U2U services on terrorism, CSEA and other duties.

## Our reasoning

### How this measure works

2.130 Content moderation systems and processes typically rely on a service's content policies, which form the basis for content moderation practices.

2.131 Content policies often exist in two forms – external and internal.

- External content policies are publicly available documents aimed at users of the service which provide an overview of a service provider's rules about what content is allowed and what is not. These are often in the form of terms of service or community guidelines. It is a requirement of the Act that providers have terms of service that include provisions specifying how individuals are to be protected from illegal content (for example, through moderation), and our recommendations on this are in chapter 10 in this Volume: 'Terms of service and publicly available statements'.

[85] Barnardo's response to November 2023 Consultation, p.15.

[86] Online Safety Act Network (OSA Network) response to November 2023 Illegal Harms Consultation, p.44.

[87] Canadian Center for Child Protection response to May 2024 Consultation on Protecting Children from Harms Online, pp.21-22; Children's Commissioner for England response to the May 2024 Consultation on Protecting Children from Harms Online, pp.59-60; Jamie Dean response to May 2024 Consultation on Protecting Children from Harms Online, pp.14-15; UK Safer Internet Centre (UKSIC) response to May 2024 Consultation on Protecting Children from Harms Online, p.36.

- Internal content policies are usually more detailed versions of external content policies and may set out rules, standards, or guidelines (including around what content is allowed and what is not) as well as providing a framework for how policies should be operationalised and enforced.[88]

2.132 We recommend that providers of multi-risk U2U services and all providers of large U2U services should set and record internal content policies.

2.133 We do not recommend that internal content policies should be made available to users.[89] We agree with Snap that the publication of internal content policies to users risks helping perpetrators to evade content moderation.[90]

2.134 In setting and recording internal content policies, the measure specifies that providers should have regard to their illegal content risk assessments and have processes in place for updating these policies in response to evidence of new and increasing illegal harm on their services (as tracked in accordance with Measure ICU A5 in Volume 1: chapter 5: 'Governance and accountability').

2.135 We are not recommending that providers should update their internal content policies every time they receive evidence of a new and increasing illegal harm. Rather, we are recommending that they should have processes in place to be able to do this where appropriate. Providers may be able to take other actions to protect users in response to evidence of new and increasing illegal harms on their services that do not require them to update their internal content policies, and the measure is drafted to account for this.

## Benefits and effectiveness

2.136 We consider that setting internal content policies is an important first step to establishing an effective content moderation system. In addition to the support we received for the measure in the November 2023 Consultation, several providers have separately stated

[88] Alan Turing Institute, 2021. Understanding online hate: VSP Regulation and the broader context [accessed 24 November 2024]; Meta, 2021. What's Allowed on Our Platforms? Find Out in Episode 2 of Video Series, Let Me Explain. [accessed 24 November 2024]; Trust and Safety Professional Association, no date. Policy Development. [accessed 24 November 2024]; Khoury College at Northeastern University, no date. Content Moderation Techniques. [accessed 24 November 2024]; Twitter, no date. Our approach to policy development and enforcement philosophy. [accessed 24 November 2024].

[89] Electronic Frontier Foundation response to November 2023 Consultation, p.9. We note Big Brother Watch made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.37.

[90] Snap response to November 2023 Consultation, p.10. We also note evidence from the Alan Turing Institute that publishing information on internal policies may be used by users to circumvent content moderation systems and processes. Alan Turing Institute, 2021. Understanding online hate: VSP Regulation and the broader context. pg.90 [accessed 24 November 2024].

publicly that content moderation policies play an important role in keeping users safe online.[91] [92] [93]

2.137 There is a strong argument that, at least for services that are large or multi-risk (where we are confident that providers are likely to need to moderate diverse content and may need to moderate large volumes), internal content policies establish clear guidelines for applying rules in a consistent, accurate and timely way. Clear internal content policies will allow content moderation teams to make quicker and more accurate decisions than they otherwise would. This will increase the speed with which illegal content is identified and dealt with appropriately and will reduce both the risk of illegal content being 'left up' in error and the risk of lawful content being taken down in error. The recommendation to have clear internal content policies will therefore deliver significant benefits.

2.138 We also consider there to be significant benefits in recommending that providers have regard to their illegal content risk assessments when setting and recording their policies. It is reasonable to infer that the data provided in risk assessments, about the challenges providers face, would enable providers to make higher quality decisions about what to include in their internal content moderation policies (tailored to the specific needs of their services). It is reasonable to expect that service providers that have identified illegal harms as high risk on their service could cover these harms in more detail in their internal policies. For example, fraud is a risk on many services, but it takes different forms on different services. A dating site is more likely to need to think about romance scams. An investments site is more likely to need to think about financial services offences.

2.139 Further, the Act requires that providers carry out a risk assessment when they make 'significant changes' to their services. Providers may update their internal content policies in response to these potential increases in risks of certain harms created by these changes.

2.140 We are not recommending at this time Glitch's suggestion that we should be more prescriptive about providers' internal content policies, including the recommendation that internal content policies should be gender-sensitive.[94] We consider that our recommendation that providers have regard to their risk assessment will ensure that content policies are appropriately gender sensitive. One implication of this recommendation is that where a provider's risk assessment identifies a material risk of gendered harms taking place on the service it operates, we would expect its policies to be

[91] Born Free Foundation response to November 2023 Consultation, p.5; Cats Protection response to November 2023 Consultation, p.10; Global Partners Digital response to November 2023 Consultation, p.13; Meta response to November 2023 Consultation, p.22; [✂]; Ukie response to the May 2024 Consultation on Protecting Children from Harms Online, p.12.

[92] YouTube, 2019. The Four Rs of Responsibility, Part 1: Removing harmful content. [accessed 25 November 2024]. Meta, 2020. Facebook's response to Australian Government Consultation on a new Online Safety Act. [accessed 25 November 2024]; TikTok, 2020. Creating Policies for Tomorrow's Content Platforms. [access 25 November 2024]; Mid-Sized Platform Group, 2022. Mid-Sized Platform Group – Online Safety Bill Recommendations. [accessed25 November 2024]; Twitter, no date. The Twitter Rules. [accessed 25 November 2024].

[93] In response to the 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation, several civil society organisations also recommended that providers establish and enforce comprehensive internal content moderation policies. Samaritans' response to 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation; Samaritans, 2023. Online Harms guidelines [accessed 24 November 2024]; Carnegie response to 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation.

[94] Glitch response to November 2023 Consultation, p.6. We note VAWG Alliance made a similar point response to the May 2024 Consultation on Protecting Children from Harms Online, p.12.

crafted in such a way as to allow for the effective moderation of illegal content relating to such harms. In February 2025, we will publish our draft guidance on protecting women and girls – and on assessing and reducing the risk of harm to them specifically – which providers may reference to further improve the gender sensitivity of their internal content policies.

2.141 We also consider that there are significant benefits of providers having processes in place for updating internal content policies in response to evidence of new or increasing illegal harm on the service. Where there are systems and processes in place to ensure policies are updated, these should improve the quality of these policies, and by extension improve the performance of providers' content moderation systems and better protect users from harm.

## Costs and risks

2.142 Service providers that do not currently have internal content policies will incur the costs of developing them. Some service providers may choose to use external experts, which could increase costs. Approving new policies may also take up senior management's time, which would add to the upfront costs. Since our November 2023 Consultation, we have further analysed these costs for the purposes of our May 2024 Consultation. We estimated that the cost to providers of smaller U2U services of implementing the equivalent measure in Children's Safety Codes could be in the region of £3,000 to £7,000.[95] While this cost estimate relates to developing an internal content policy relating to content harmful to children, we expect that the costs of developing such a policy relating to illegal harms could be similar for many smaller providers. This is because the development process, staff involvement and time required is likely to be similar. For both illegal harm and content harmful to children, costs are likely to differ between providers depending on the type and number of harms present on the service.

2.143 Providers of large services may require more complex content policies, as the way in which harm can materialise is likely to be more varied on such services and the governance requirements needed to implement them are also likely to be more complex. These factors may increase costs due to the increased amount of time required to design more complex policies. These costs could reach the tens of thousands or more.[96] In addition, there may be some small ongoing costs to ensure these policies remain up to date over time (for example, to take into account new and increasing illegal harms).

2.144 Some service providers will already have policies in place which at least partly address this measure. For these service providers, the proposed measure will mainly involve costs to update existing policies in line with risk assessments and any emerging evidence of harms.

2.145 These costs are mitigated by the flexibility of the measure, as we have set out high-level recommendations that give providers flexibility over how they choose to implement them.

[95] Assuming a service required three weeks of time across professional occupations (legal/regulatory staff) and four hours of senior management time to develop an internal content moderation policy. Based on our wage estimate assumptions as set out in Annex 5.

[96] These cost estimates do not change the approach on which we consulted in our November 2023 Consultation, but add further detail to support our position.

## Rights impact

### Freedom of expression and freedom of association

2.146 We consider that this measure has the potential to impact on users' rights to freedom of expression for the reasons set out in relation to the measures on reviewing and assessing content and swiftly taking down content, since it would inform providers' decisions made according to that measure.

2.147 In addition to the impacts identified in the measures on reviewing, assessing and swiftly taking down content, we are of the view that this measure has the potential to interfere with users' rights to freedom of expression if internal content moderation policies define the content in scope of these policies more widely than is necessary to comply with the Act. However, nothing in this measure requires or encourages providers to do this. As a matter of their own right to freedom of expression, providers are entitled to decide what content they want to allow on their service, so long as they protect UK users from the types of harmful content (including illegal content) regulated by the Act.

2.148 We consider there may also be positive impacts on users' right to freedom of expression and freedom of association from providers implementing this measure. Internal content moderation policies can set out a level of detail that may not be practical to do in external facing policies, providing content moderators with greater clarity on the type of content that is illegal content and priority illegal content, resulting in a higher degree of content being identified appropriately. Where they are likely to be dealing with large volumes of content, the process of considering these matters in advance and preparing a policy would tend to improve internal scrutiny, and improve the consistency and predictability of decisions, in a way which we consider would also tend to protect users' rights.

2.149 We therefore consider that the impact of this measure on users' rights to freedom of expression and freedom of association, above and beyond the requirements of the Act, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their safety duties about illegal harms under the Act. Taking this, and the benefits to users and affected persons into consideration, we consider that any impact on rights of freedom of expression and association from this measure is proportionate.

### Privacy

2.150 We consider that this measure has the potential to impact on users' right to privacy to the extent that a service provider's internal policies describe or define content relevant to their safety duties by reference to information in relation to which a user would have a reasonable expectation of privacy, or by reference to personal data.

2.151 However, where service providers are likely to be dealing with large volumes of content, the process of considering these matters in advance and preparing a policy would be likely to improve internal scrutiny, and improve the consistency and predictability of decisions, in a way which we consider would also be likely to protect users' privacy. Taking this, and the benefits to users and affected persons into consideration, we consider that any impact on privacy rights from this measure is proportionate.

### Data Protection

2.152 Providers are required to comply with applicable data protection laws including when implementing safety measures.[97] Having a set of policies in place will also encourage consistency and predictability in content moderation, which will help to secure that any processing of personal information is appropriate.

2.153 We therefore consider that (assuming service providers comply with applicable data protection laws) this measure is likely to constitute the minimum degree of interference required to secure that service providers fulfil their safety duties about illegal content under the Act. Taking this, and the benefits to users and affected persons into consideration, we consider that any impact on data protection rights from this measure is proportionate.

## Who this measure applies to

2.154 We expect that the benefits of applying this measure to providers of multi-risk services will be substantial, given the risks that providers of these services will have identified in their risk assessments. Our analysis suggests that internal content policies are an important part of an effective content moderation system that reduces harms to end users. We consider that services in scope of this measure are unlikely to be able to moderate content effectively without such policies. As the costs of this measure are likely to be relatively small for many service providers, we consider it proportionate to apply it to all providers of multi-risk services.

2.155 This measure may have fewer benefits for providers of large services with low risks of illegal harm since there may be less scope for reducing harm to users from illegal content. However, as explained in 'Our approach to developing Codes measures' we consider that applying this and other measures to such services will have benefits for users as these services have the potential to affect many users and the nature of illegal content can change over time. In particular, we note that providers of large services may still have substantial volumes of content to moderate and a large number of content moderators. We consider that the measure will promote consistency of approach in this situation. Providers of large services are also likely to have sufficient resources to develop or adjust policies in line with the measure.

2.156 As set out in paragraphs 2.125 and 2.126 , several respondents suggested that this measure should also apply to all providers of single-risk services or to providers of all services (including smaller low-risk ones).[98] [99] We consider that the benefits of having an internal content policy are likely to be materially lower for smaller services that are low-risk for all types of harms as such services will not need to review very much (if any) potentially illegal content. Possible examples of such services might include those where the U2U component of a service is a peripheral part of the main service (perhaps including some food delivery companies). We therefore remain of the view, based on the evidence available to us at

[97] In determining what this requires of them, they should have regard to any relevant guidance from the ICO.

[98] Age Verification Providers Association response to November 2023 Consultation, p.2; NSPCC response to November 2023 Consultation, p.20; VerifyMy response to November 2023 Consultation, p.6; Yoti response to November 2023 Consultation, p.15.

[99] Barnardo's response to November 2023 Consultation, p.15; OSA Network response to November 2023 Consultation, p.44. We note that C3P (p.21-22), Children's Commissioner for England (pp.59-60), James Dean (pp.14-15) and UKSIC (p.36) made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online.

present, that it would not be proportionate to extend this measure to such service providers.

2.157 We note the arguments that this measure should apply to some or all providers of single-risk services. As explained in ‘Our approach to developing Codes measures’, we expect to consult again on this in Spring 2025. We do not consider it appropriate to delay the start date of the regulatory regime to accommodate a further consultation on this policy point. As outlined in paragraphs 2.154, there is benefit to the current scope of this measure.

2.158 We are therefore recommending this measure for all providers of large U2U services and all providers of multi-risk services.

## Conclusion

2.159 The analysis above shows that the measure we are recommending is likely to deliver significant protections for users from harm and that the costs and impact on rights that will result from it are proportionate. We have made a slight amendment to the measure, to clarify to providers that in setting and recording internal content policies, they should have processes in place for updating these policies in response to evidence of new and increasing illegal harms on the service (as tracked in accordance with the Measure ICU A5 in Volume 1: chapter 5: ‘Governance and accountability’). With the exception of this slight amendment, we have decided to leave the measure largely unchanged from the amended measure we proposed in our May 2024 Consultation.

2.160 All providers of multi-risk and large services should set and record internal content policies, and, in doing so, have regard to their illegal content risk assessments and have processes in place for updating these policies in response to evidence of new and increasing illegal harm on the service (as tracked in accordance with Measure ICU A5 in Volume 1: chapter 5: ‘Governance and accountability’). We consider that internal content policies establish clear guidelines for applying rules in a consistent, accurate, and timely way.

2.161 This measure will be included in our Codes of Practice for U2U services on terrorism, CSEA and other duties. It is referred to within these Codes as ICU C3.

# Measure on performance targets

2.162 In our November 2023 Consultation, we proposed that service providers should set performance targets for their content moderation functions and measure whether they are achieving them. We proposed that this measure should apply to all providers of large U2U services and all providers of multi-risk U2U services.

2.163 We considered that if providers are clear about the content moderation outcomes they are trying to achieve and are measuring whether they are achieving them, they will be better able to plan how to configure their systems to meet these goals and optimise the operation of these systems.

## Summary of stakeholder feedback[100]

2.164 In addition to those stakeholders who expressed broader support for the full package of content moderation measures outlined in paragraph 2.14, [✂].[101]

2.165 There were several areas where stakeholders felt this measure should be amended. These included (but were not limited to):

- the flexibility of the measure;
- how we determine that a provider has complied with the measure;
- targets for terms of service versus illegal content;
- concerns on the practical applicability of time targets;
- the definition of accuracy in the measure;
- balancing the desirability of moderating content quickly as well as accurately;
- arguments against us recommending performance targets for content moderation;
- concerns about unintended incentives created by performance targets;
- privacy rights; and
- who the measure applies to.

2.166 We outline these stakeholder concerns in more detail below, and address additional stakeholder responses in Annex 1.

### The flexibility of the measure

2.167 Several stakeholders argued that providers should have the flexibility to set their own performance targets, and that we should not recommend specific types of targets.[102] Some providers shared evidence about targets they already use and consider to be more effective than those listed in our proposed measure, including targets related to reducing the number of users exposed to harm.[103] For example, Google described Violative View Rate (VVR) as its “North Star” for content moderation on YouTube.[104]

[100] Note this list in not exhaustive, and further responses can be found in Annex 1.

[101] [✂].

[102] Center for Data Innovation response to November 2023 Consultation, p.10; Google response to November 2023 Consultation, p.35; Mid Size Platform Group response to May 2024 Consultation on Protecting Children from Harms Online, p.10; Pinterest response to November 2023 Illegal Harms Consultation, p.7. We note that Google (p.26) and Pinterest (p.15) made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online.

[103] Google response to November 2023 Consultation, p.35; Pinterest response to November 2023 Consultation, p.7. We note that Pinterest made a similar point in response to May 2024 Consultation on Protecting Children from Harms Online, p.14.

[104] Google response to November 2023 Consultation, p.35.

2.168 In contrast, several stakeholders argued that the measure should be more prescriptive. Suggestions included:

- recommending baselines for performance targets (including setting a 24-hour target for the take-down of CSAM or time limits for the time it takes to review reports of fraudulent adverts from trusted flaggers;[105] [106]
- that we recommend consistent performance metrics that providers should use to allow cross-industry comparison;[107]
- that we set expectations on what performance targets should be on different types of services;[108] and
- that we recommend the outcomes performance targets should achieve.[109]

2.169 The New Zealand Classification Office made suggestions about types of performance targets providers should not use; for example, the amount of content removed, or amounts of referrals or requests actioned by the service.[110]

2.170 Some stakeholders raised concerns about providers skewing metrics to optimise the appearance of their performance if given the flexibility to set their own performance targets.[111]

2.171 We address these points in paragraphs 2.205 to 2.207 in the section entitled 'How this measure works.'

## How we determine that a provider has complied with the measure

2.172 Microsoft requested clarification about how we would determine when a provider had failed to set satisfactory performance targets, and the factors, data, or documents we would expect a provider to cite when setting targets.[112]

2.173 We address this point in paragraph 2.203 in the section entitled 'How this measure works'.

## Concerns on the practical applicability of time targets

2.174 Some stakeholders questioned the applicability of time-based performance targets to different types of content. [✂] and techUK raised concerns that timelines to make moderation decisions may vary depending on the type of content involved. [113] Meta argued that asking providers to set specific response times for the moderation of content does not account for the nuance in assessing cases with differing levels of complexity. It argued that even when violating content is part of the same 'category' of violation, no two violations

[105] Marie Collins Foundation response to November 2023 Illegal Harms Consultation, p.9.
[106] Lloyds Banking Group response to November 2023 Illegal Harms Consultation, p.5.
[107] Institute for Strategic Dialogue response to November 2023 Illegal Harms Consultation, p.9
[108] Refuge response to November 2023 Consultation, p.12
[109] 5Rights Foundation response to November 2023 Illegal Harms Consultation, p.21.
[110] New Zealand Classification Office response to the November 2023 Consultation, p.7
[111] 5Rights Foundation response to November 2023 Consultation, p.21; Institute for Strategic Dialogue response to November 2023 Consultation, p.9.
[112] Microsoft response to the November 2023 Consultation, p.10.
[113] [✂]; techUK response to November 2023 Consultation, p.8.

are the same, making it impractical to set single turnaround times for a moderation system as a whole. [114]

2.175 Google requested clarification of what stage of the content moderation process providers are recommended to set time targets for. It suggested that the measure, as drafted in the November 2023 Consultation, could be interpreted to be recommending providers to set time targets for all illegal content, even if it had not been reported.[115] It raised concerns that this would imply we are recommending providers undertake general monitoring.[116]

2.176 We address these points in paragraph 2.211 and 2.212 in the section entitled 'How this measure works'.

2.177 Meta said that this measure conflicts with the measure on a policy of prioritisation for review, as when prioritisation applies when content would be reviewed would be in a fluid state dependent on the prioritisation queue.

2.178 We address this point in paragraph 2.225 in the section entitled 'How this measure works'.

### The definition of accuracy in the measure

2.179 Meta, as well as the Centre for Competition Policy, requested more information on how we define accuracy of decision-making in the measure.[117] Both respondents suggested that information about complaints, and whether they were successfully appealed, might be part of this definition.[118]

2.180 We address these points in paragraph 2.214 in the section entitled 'How this measure works'.

### Balancing the desirability of moderating content quickly as well as accurately

2.181 An individual and the Centre for Competition Policy expressed support for our recommendation that providers balance the desirability of taking illegal content down swiftly with the desirability of making accurate moderation decisions.[119]

2.182 Open Rights Group argued that we did not provide enough guidance on what this balance should look like, and that providers are incentivised to prioritise speed.[120]

2.183 Pinterest suggested that performance targets for making decisions quickly as well as accurately can be in tension with each other, particularly where a small number of edge cases can skew average turnaround times due to the additional analysis required to make an accurate decision.[121] Snap suggested that the correct balance between the timeliness and accuracy of decision-making can change in response to external events, new risks and

[114] Meta response to the November 2023 Consultation, p.22. We note that Meta made a similar point in the May 2024 Consultation on Protecting Children from Harms Online, p.21-22.

[115] Google response to November 2023 Illegal Harms Consultation, pp.36-37.

[116] Google response to November 2023 Consultation, pp.36-37.

[117] Centre for Competition Policy response to November 2023 Illegal Harms Consultation, p.16; Meta response to November 2023 Consultation, p.22.

[118] Centre for Competition Policy response to November 2023 Consultation, p.16; Meta response to November 2023 Consultation, p.23.

[119] Are, C. response to November 2023 Consultation, p.7; Centre for Competition Policy response to November 2023 Consultation, p.19.

[120] Open Rights Group response to November 2023 Consultation, p.2.

[121] Pinterest response to November 2023 Consultation, p.7. We note that Pinterest made a similar point in response to the May 2024 Consultation, p.15.

the evolution of a service and its user base.[122] It therefore suggested the importance of quality assuring performance targets to ensure they remain effective.[123]

2.184 We address these points in paragraphs 2.217 and 2.218 in the section entitled 'How this measure works'.

### Targets for terms of service versus illegal content

2.185 Google raised concerns that the measure requires providers to produce separate performance targets for moderating illegal content to content that breached a provider's terms of service.[124] Match Group agreed it was important to set and measure delivery against performance targets, while cautioning against setting requirements that may incentivise providers to set a higher threshold for content to be taken down, based on illegality rather than if it was just deemed harmful by a provider.[125]

2.186 We address these points in paragraphs 2.220 and 2.221 in the section entitled 'How this measure works'.

### Arguments against us recommending performance targets for content moderation

2.187 Meta proposed that we should make a general recommendation that content should be reviewed swiftly, rather than recommending that providers set performance targets.[126] It suggested that this approach would align with other regulatory regimes such as the Digital Services Act (DSA).[127]

2.188 While Pinterest did not object to our recommendation to set performance targets, both Pinterest and Mid Size Platform Group argued against these targets being used to determine the overall effectiveness of content moderation.[128] Pinterest argued that, in particular, turnaround time should not be treated as a determinative factor in whether providers' content moderation systems are effective.[129]

2.189 We address these points in paragraph 2.228 and 2.229 in the section entitled 'Benefits and effectiveness'.

### Unintended incentives created by performance targets

2.190 Several stakeholders raised concerns that time targets would incentivise providers to remove too much content on their services to appear to be performing better against these targets.[130]

[122] Snap response to November 2023 Consultation, p.10.

[123] Snap response to November 2023 Consultation, p.10.

[124] Google response to November 2023 Consultation, p.35.

[125] Match Group response to November 2023 Consultation, pp. 9-10.

[126] Meta response to November 2023 Consultation, p.23. We note that Meta made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.21.

[127] Meta response to November 2023 Consultation, p.23.

[128] Mid Size Platform Group response to November 2023 Consultation, p.8; Pinterest response to November 2023 Consultation, p.7. We note that Pinterest made a similar point in response to the May 2024 Consultation, p.15.

[129] Pinterest response to November 2023 Consultation, p.6.

[130]Big Brother Watch response to November 2023 Consultation, pp.4-5; Electronic Frontier Foundation response to November 2023 Illegal Harms Consultation, p.8; Meta response to November 2023 Consultation, p.22; Mid Size Platform Group response to November 2023 Consultation, p.8; Pinterest response to November

2.191 Pinterest also suggested that providers may be incentivised to reject valid appeals and conduct less careful reviews of nuanced content, so as to achieve better accuracy rates against targets.[131]

2.192 We address these points in paragraph 2.232 and 2.234 in the section entitled 'Benefits and effectiveness'.

2.193 Some stakeholders raised concerns about the implications of such incentives on users' freedom of expression rights.[132]

2.194 We address this point in paragraph 2.240 to 2.244 in the section entitled 'Rights impact'.

## Privacy rights

2.195 The ICO suggested that we should include reference to requirements in data protection law to ensure personal information is accurate when providers are balancing the desirability of taking illegal content down swiftly against the desirability of making accurate moderation decisions.[133] It argued that this is particularly important where CSEA content is detected because of the risk of incorrect reporting to the NCA.[134]

2.196 We address this point in paragraphs 2.245 – 2.247 in the section entitled 'Rights impact'.

## Who this measure applies to

2.197 The Online Dating and Discovery Association said performance targets for content moderation for a team of one to three people would look very different than for a large service provider with a trust and safety department.[135]

2.198 Some stakeholders argued that the measure should apply to all providers of services with a specific risk (as well as multi-risk and large providers).[136]

2.199 We note that some stakeholders argued that an equivalent measure proposed in the May 2024 Consultation should apply to all service providers.[137]

2.200 We address these points in paragraphs 2.250 to 2.253 in the section entitled 'Who this measure applies to'.

2023 Consultation, p.7; Reddit response to November 2023 Illegal Harms Consultation, p. 26. We note that Pinterest (p.15) and Meta (pp.21-22) made a similar point in response to the May 2024 Consultation.

[131] Pinterest response to November 2023 Consultation, p.7.

[132] Big Brother Watch response to November 2023 Consultation, pp.4-5; Electronic Frontier Foundation response to November 2023 Consultation, p.8; Meta response to November 2023 Consultation, p.22; Reddit response to November 2023 Consultation, p. 26. We note that Big Brother Watch response to the May 2024 Consultation on Protecting Children from Harms Online, p.38-39.

[133] ICO response to November 2023 Consultation, p.15.

[134] ICO response to the November 2023 Consultation, p.15.

[135] Online Dating and Discovery Association response to November 2023 Consultation, p.2.

[136]Age Verification Providers Association response to November 2023 Illegal Harms Consultation, p.2; VerifyMy response to November 2023 Consultation, p.6; Yoti response to November 2023 Consultation, p.16.

[137] C3P response to the May 2024 Consultation, pp.21-22; Children's Commissioner for England response to May 2024 Consultation, pp.59-60; Jamie Dean response to May 2024 Consultation, pp.14-15; UKSIC response to May 2024 Consultation, p.36.

## Our decision

2.201 We have decided to recommend the measure broadly as we proposed in the November 2023 Consultation. We have made some small amendments to some of the components of the measure discussed in this section:

- To clarify when time targets should apply, the measure now recommends that providers should set **targets for the time period for taking relevant moderation action**. For most providers, this is the time it takes for a provider to review, assess and take down content from when it has reason to suspect that content may be illegal content.[138] For providers for whom takedown is currently technically infeasible, we consider the relevant moderation action is reviewing and assessing at least suspected CSEA or proscribed organisation content.[139]
- The measure now says that providers may set performance targets either for illegal content or (if they make judgments against their own terms of service) an **illegal content proxy.** This is to give providers the choice to set performance targets for content that is illegal or content that violates their terms of service, aligned with the choice they make for measure ICU C1.
- The measure also now says that providers should balance the **need** to take relevant moderation action swiftly with the **importance** of making accurate moderation decisions. These words replace the term **desirability** which we used in the November 2023 Consultation. This is to clarify our expectation that providers must balance speed and accuracy to set appropriate performance targets for their services.

2.202 The full text of the measure can be found in our Illegal Content Codes of Practice for U2U services and is referred to within these as ICU C4. This measure will be included in our Illegal Content Codes of Practice for U2U services for terrorism, CSEA and other duties.

## Our reasoning

### How this measure works

2.203 We recommend that providers of services which are large or multi-risk should set performance targets for their content moderation functions and track whether they are meeting these. We do not consider it necessary to provide more detail of the process which providers should adopt when setting these targets, as per the queries from Microsoft, except for our recommendation that they should balance the need to take relevant content moderation swiftly with the importance of making accurate moderation decisions.[140]

2.204 We recommend that, at a minimum, providers' performance targets should include targets for the time period for taking relevant content moderation action, and targets relating to the accuracy of such content moderation decisions.

2.205 Several stakeholders suggested ways in which we should be more prescriptive about the specific performance targets providers should set and we recognise the concern that

[138] In line with our recommendations in ICU C1.3 and ICU C2.3 in the Illegal Content Codes of Practice for U2U services.

[139] See paragraphs 2.47 to 2.57 above, where we set out our reasons for our approach.

[140] Microsoft response to November 2023 Consultation, p.10.

providers will seek to set metrics which make them look good.[141] We are concerned that being more prescriptive at this early stage in the regulatory regime – with the limited information available to us – could have significant unintended consequences and may make users less safe. We do not currently have evidence that would enable us to specify in detail how providers should configure their content moderation systems, including the baseline at which providers should set their performance targets, nor to restrict providers from setting specific types of performance targets. We do not consider a one size fits all approach to be appropriate due to the wide range of services this measure applies to, and do not currently have enough evidence to specify what differing expectations for targets should be on different types of services. We also consider that providers need the flexibility to adjust performance targets to suit the needs of their service and adjust them over time as circumstances for their service change.

2.206 While we do not recommend the outcomes performance targets should achieve in this measure, we specify in the measure on resourcing that providers should resource their content moderation functions to give effect to the performance targets that we recommend they set. We consider that the combined outcome of this measure and the measure on resourcing is that providers' content moderation functions should be sufficiently resourced to meet performance targets.

2.207 In contrast, other stakeholders (mainly service providers) suggested that we offer providers complete flexibility, and do not recommend any types of targets that providers should include in their suite of performance targets.[142] However, we consider it necessary and proportionate to recommend some targets which, at a minimum, providers should include in their performance targets. We consider that providers should at least be setting performance targets for the time period for taking relevant content moderation action, as well as the accuracy of decision-making.

### Targets for the time period for taking relevant content moderation action

2.208 We recommend that providers' performance targets include targets for the time period for taking relevant content moderation action. For most providers, relevant content moderation action means the steps outlined in the measures on reviewing, assessing and swiftly taking down content.[143] For providers for whom takedown is currently technically infeasible, we consider the relevant moderation action is reviewing and assessing at least suspected CSEA or proscribed organisation content.

2.209 When a provider has reason to suspect content may be illegal content, it should review the content and either:

- make an illegal content judgement in relation to the content; or

[141] 5Rights Foundation response to November 2023 Consultation, p.21; Institute for Strategic Dialogue response to November 2023 Consultation, p.9; Lloyds Banking Group response to November 2023 Consultation, p.5; Marie Collins Foundation response to November 2023 Consultation, p.9; Refuge response to November 2023 Consultation, p.12; New Zealand Classification Office response to November 2023 Consultation, p.7.

[142] Center for Data Innovation response to November 2023 Consultation, p.10; Google response to November 2023 Consultation, p.35; Pinterest response to November 2023 Consultation, p.7. We note that Pinterest (p.15), Mid Size Platform Group (p.10), and Google (p.26) made a similar point in response to the May 2024 Consultation.

[143] We note that relevant moderation actions will be different for providers for whom it is currently not technically feasible to take down content.

- where the provider is satisfied that its terms of service prohibit the types of illegal content which it has reason to suspect exist, consider whether the content is in breach of those terms of service.

2.210 If a provider determines that content is illegal content, or is in breach of its terms of service, it should swiftly take the content down unless it is not currently technically feasible for it to do so.

2.211 We note Google's concerns that this measure, as drafted in the November 2023 Consultation, could be interpreted as recommending that providers set performance targets which require them to proactively identify content.[144] We do not consider that a measure that recommended that providers set such a target would be appropriate, as we do not have the evidence to recommend that providers should set targets to proactively identify content for review.[145] We have therefore amended the measure to be clear that providers should set time targets from the point at which they have reason to suspect content is illegal.[146]

2.212 We agree with stakeholder arguments that different types of content may require different review timelines.[147] We consider our measure accounts for this by giving providers the flexibility to set different time targets for different content as appropriate for their service.

### Targets for the accuracy of decision-making

2.213 We also recommend that providers' performance targets include targets for the accuracy of decision-making. For example, we understand that some providers do this by tracking the rate of appeals as a measure of the accuracy of the decisions that are taken.[148]

2.214 We do not consider it necessary to be more prescriptive in the definition of accuracy of decision-making in the measure.[149] We consider that providers are best placed to set appropriate performance targets for accuracy based on what is most suitable for their services, including the extent to which they use information about complaints and appeals.

### Other targets

2.215 We understand that many larger providers use metrics and targets for reducing users' exposure to harmful content for user safety.[150] These outcomes-based metrics have the

[144] Google response to November 2023 Consultation, pp.36-37.

[145] As outlined in chapter 4 of this Volume: 'Automated content moderation'.

[146] We explain some of the ways in which a provider may become aware of content it suspects to be illegal content in paragraph 2.52.

[147] Meta response to November 2023 Consultation, p.22; [✂]; techUK response to November 2023 Consultation, p.8. We note that Meta made a similar point in the May 2024 Consultation on Protecting Children from Harms Online, p.21-22.

[148] Twitch,2022. H2 2022 Transparency Report (twitch.tv) [accessed 25 November 2024]; Pinterest, 2023. Digital Services Act Transparency Report | Pinterest Policy. [accessed 25 November 2024].

[149] Centre for Competition Policy response to November 2023 Consultation, p.16; Meta response to the 2023 Consultation, p.23.

[150] Meta described metrics reflecting the viewing of violative content, before the content was actioned, as "the number we hold ourselves accountable to". Facebook, 2018., Understanding the Facebook Community Standards Enforcement Report. [accessed 25 November 2024]; YouTube described these metrics as "the primary metric [we use] to measure our responsibility work". YouTube, 2021. Building greater transparency and accountability with the Violative View Rate. [accessed 25 November 2024]. We note that Google shared similar evidence in its response to the November 2023 Illegal Harms Consultation, p.35 and that Pinterest shared similar evidence in its response to the November 2023 Illegal Harms Consultation, pp.6-7 and in its response to the May 2024 Consultation on Protecting Children from Harms Online, p.14.

advantage of capturing not just content moderation, but also other factors that can affect the extent of users' exposure to harm on a service. This includes the functionalities, features, and design of the service that may have a large impact on the extent of users' exposure to illegal content. However, we note that there may be an impact on human rights depending on how this target is achieved.

2.216 While we welcome providers using exposure targets for this wider purpose, these do not capture all aspects of protecting users from illegal content and are not specific to content moderation. In particular, such targets are less relevant for harms that are targeted at a particular individual (such as harassment), and they do not capture the extent to which a particular individual has been subjected to such harms by illegal content.

2.217 We consider targets that capture speed and accuracy of content moderation specifically are important, and particularly help with harms where wide exposure is less relevant. We have therefore decided to recommend these two targets specifically for content moderation.

2.218 We do not consider there to be any tension between providers setting time and accuracy targets for content moderation and also having other wider performance metrics for which they set targets. This could include exposure metrics that relate to the overall performance of a service in addressing harms. We would welcome providers wanting to design a range of targets related to user safety that are appropriate to the risks on their services and decision-making processes that go beyond the types of performance targets listed in this measure.

### Balancing the need to take relevant moderation action swiftly with the importance of making accurate moderation decisions

2.219 In this measure, we recommend that in setting its targets, the provider should balance the need to take relevant moderation action swiftly with the importance of making accurate moderation decisions.

2.220 We have replaced the term **desirability** with the words **need** and **importance** in this part of the measure to clarify our expectation that speed and accuracy are not only desirable, but are essential components of an effective content moderation system. Providers should set their performance targets in a way that pursues both speed and accuracy of moderation and does not solely pursue one of these factors to the detriment of the other. We consider that the tension between these two factors highlighted by Pinterest is a beneficial feature of this measure and incentivises providers to strike a balance between these factors, which makes their performance targets most effective at protecting users on their service.[151] We do not consider that it is necessary for us to provide further guidance on how providers should achieve this balance when setting performance targets, as recommended by Open Rights Group.[152] Providers will be best placed to choose how to balance these factors in a way that is suitable for their services. However, while we are not being prescriptive, we would expect providers to be able to justify why they have set the performance targets it has, including how they have balanced speed and accuracy when making this decision and why the targets are reasonable.

2.221 We note Snap's point that the quality assurance of performance targets may help to address the fact that the correct balance between the timeliness and accuracy of decision-

[151] Pinterest response to November 2023 Consultation, p.7. We note that Pinterest made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.15.

[152] Open Rights Group response to November 2023 Consultation, p.2.

making can change over time and in response to external events or new risks.[153] In our measure on resourcing, we recommend that providers resource their content moderation functions to give effect to their performance targets, having regard to the propensity for external events to lead to a significant increase in demand for content moderation on the service. We expect that service providers will build flexibility into the resourcing of their content moderation function, so that where an external event or new risk on a service causes a change in the levels of demand for content moderation, such demands can be met without the need to adjust performance targets. However, the recommended measures leave providers with the ability to update performance targets where appropriate.

### How the measure fits with other Content Moderation measures

2.222 We agree with Google's comment that the drafting of the measure proposed in our November 2023 Consultation could be understood as requiring providers to set separate performance targets for illegal content and content that was in breach of a provider's terms of service. [154] We also agree with the point made by Match Group about the risk of unintended consequences occurring from setting targets for illegal content rather than more general harms prohibited on the service.[155]

2.223 Requiring providers to set separate performance targets for illegal content only and not allowing them to set targets for violations of their terms of service was not the original policy intent behind this measure, nor would it be compatible with the amendment we have made to time targets (as described in paragraph 2.211).

2.224 We have therefore amended the measure to give providers flexibility to set performance targets for illegal content or illegal content proxy.[156] The measure now recommends that providers should have targets for the period for taking relevant moderation action. Relevant moderation action for most kinds of service involves a provider reviewing content it has reason to suspect may be illegal content and then either:

- the provider should make an illegal content judgement in relation to the content and determine if the content is illegal content. If it determines the content is illegal content, it should swiftly take it down; or
- where the provider is satisfied that its terms of service prohibit the types of illegal content which it has reason to suspect exist, it should consider whether the content is in breach of those terms of service. If it considers that content is in breach of its terms of service, it should swiftly take it down.[157]

2.225 Meta argued that this measure is not compatible with the measure on prioritisation.[158] We disagree, as we consider this measure gives providers the flexibility to set performance

[153] Snap response to November 2023 Consultation, p.10.
[154] Google response to November 2023 Consultation, p.35
[155] Match Group response to November 2023 Consultation, pp.9- 10.
[156] In the Codes, we define "illegal content proxy" as content that a provider determines to be in breach of its terms of service, where: the provider had 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.
[157] We note that relevant moderation actions will be different for providers for whom it is currently not technically feasible to take down content.
[158] Meta response to November 2023 Consultation, p.22.

targets at a level which allows them to prioritise content for review as appropriate for their service.

## Benefits and effectiveness

2.226 Performance targets provide a quantitative target for the effectiveness of content moderation efforts. Monitoring compliance with performance targets helps evaluate the performance of content moderation systems and processes.

2.227 We understand that many services set performance targets for the operation of their content moderation functions and measure whether they are achieving these.[159]

2.228 We consider that recommending providers set performance targets – rather than a more general outcomes-based recommendation like that suggested by Meta– will have important benefits.[160] Where providers explicitly set targets and measure performance against them, they are more likely to be able to optimise the design of their moderation functions to achieve the goals underlying the targets than they would be if they did not set targets. For example, we consider that all else being equal, a provider with an overall aspiration of swiftly and accurately moderating illegal content would be more likely to do so if it set clear and explicit targets for timeliness and accuracy of moderation action than if it did not. As explained in more detail in paragraphs 2.230 to 2.233, content being moderated quickly and accurately has important benefits for user safety.

2.229 Monitoring performance against targets over time also gives providers data about how their content moderation system and processes are performing and allows them to make changes to their systems and processes to better protect users based on these data. We agree with stakeholder arguments that data on time and accuracy will not be the only relevant factors of the performance of a content moderation system but consider that these data are nonetheless a useful indication of this.[161]

### Time targets

2.230 Users are better protected from illegal content if decisions which will result in action on content are made quickly, which means there is a clear benefit to providers considering the need for swift review of harmful content when setting their performance goals for content.

[159] In response to the 2022 Illegal Harms Call for Evidence, OnlyFans told us that, within two minutes of an attempted upload, all content is triaged by automated technologies, and reviewed by human moderators in the pre-check team, and that all content that passes this initial review is then also reviewed by a human content moderator within 24 hours of being posted onto the platform. OnlyFans response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation. Twitch told us its time target for responding to live video content is within 10 minutes of receiving a report, and the median time taken to respond to such content has consistently been within that target. Ofcom/Twitch meeting, 23 March 2023. In H1 2024, it responded to 76% of reports within that time. Twitch, 2024. H1 2024 Transparency Report. [accessed 25 November 2025]. Via stakeholder engagement, a large gaming service told us its target response time is 3 hours max for high priority content, and 24 hours max for user reports and appeals. [✂]. TikTok records its removal rate within 24 hours. TikTok,2023. Community Guidelines Enforcement Report. [accessed 25 November 2024]. Snapchat records 'Turnaround Time' and publishes the medium time for various platform violations. Snapchat,2023. Transparency Report Glossary. [accessed 24 November 2024].

[160] Meta response to November 2023 Consultation, p.23. We note Meta made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.21.

[161] Mid Size Platform Group response to November 2023 Consultation, p.8; Pinterest response to November 2023 Consultation, p.7.

2.231 As set out above, a relatively small minority of services are currently technically unable to take content down. Where take down is technically infeasible, there are still benefits to providers setting time targets for the review of at least content they suspect to be CSEA or proscribed organisation content. The quicker CSEA is (correctly) reported, and the quicker a proscribed organisation account is (correctly) identified, the quicker providers will be able to report CSAM to the NCMEC or the NCA and the quicker they will be able to shut down accounts run by proscribed organisations.

2.232 We note concerns from stakeholders that time targets might result in the over-removal of content.[162] Stakeholders raised similar concerns in our 2022 Illegal Harms Call for Evidence.[163] However, we consider that this risk is mitigated by the flexibility of this measure, which allows providers to set their own time targets at a level which does not incentivise them to do this.

### Accuracy targets

2.233 We note the risk raised by some stakeholders that providers having a disproportionate focus on speed of content removal could lead to pressure on systems, poorer quality decisions and, in turn, a decrease in accuracy. However, this risk is mitigated by our recommendation that providers also set targets for the accuracy of decision-making and balance the need to take relevant content moderation action swiftly against the importance of making accurate moderation decisions.

2.234 We consider that the risk outlined by Pinterest of accuracy targets on content moderation creating an incentive for providers to reject valid appeals is mitigated by our recommendation in chapter 6 of this Volume: ‘Reporting and complaints’ that providers set and monitor their performance for the determination of relevant complaints which are appeals against performance targets.[164] We recommend that these targets should include targets for the accuracy of decision-making.

## Costs and risks

2.235 Service providers will incur one-off costs in designing and setting up suitable performance metrics and targets. This could involve one-off system changes to determine (for example) the number of views of content subsequently found to be illegal, or to track the time that has elapsed between content being reported and content being assessed or actioned if found to be violative.

2.236 To assess accuracy of content moderation decisions, service providers may take a sample of these decisions and re-assess them, which may incur significant ongoing costs. There may also be further ongoing costs, such as those associated with data storage.

[162] Big Brother Watch response to November 2023 Consultation, pp.4-5; Electronic Frontier Foundation response to November 2023 Consultation, p.8; Meta response to November 2023 Consultation, p.22; Mid Size Platform Group response to November 2023 Consultation, p.8; Pinterest response to November 2023 Consultation, p.7; Reddit response to the November 2023 Illegal Harms Consultation, p. 26. We note that Pinterest made a similar point in response to the May 2024 Consultation, p.15.

[163] As Global Partners Digital noted in its response to our Call for Evidence, “simplistic quantitative targets” such as time limits, “prioritise quantity over quality of decisions, overlook the complexity of certain cases, and prevent moderators from researching necessary context or information before making their decisions”. Global Partners Digital response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

[164] Pinterest response to November 2023 Consultation, p.7.

2.237 Since our November 2023 Consultation, we have further analysed these costs for the purposes of our May 2024 Consultation. We set out that for providers of smaller services, we expect the costs of implementing the equivalent measure in the Children's Safety Codes by creating a simple bespoke system to be approximately £8,000 to £16,000.[165] This would be where accuracy was estimated based solely on the outcome of user appeals. While this cost estimate relates to developing performance metrics and targets relating to content harmful to children, we expect that the costs of developing such a system relating to illegal harms could be similar for many smaller providers. This is because the development process, staff involvement and time required is likely to be similar. Alternatively, providers might opt to license a third-party system at a relatively low cost (such solutions are available from around £50 per month for each staff user).

2.238 For providers of large services, or those with medium or high risk of many kinds of illegal harm, the number and complexity of metrics and the associated data management processes may be significantly greater, entailing higher costs. In these cases, providers may choose to design and automate systems for proactive quality assurance of moderation decisions. As this would introduce complexity, one-off costs could run to the tens or hundreds of thousands depending on the design of the service and the volume of reports (which is likely to be linked to service size and number of risks).[166]

## Rights impact

2.239 This measure recommends that service providers in scope should set performance targets as part of their internal content policies. This measure should therefore be seen as part of a package of measures relating to content moderation for illegal content, including the measures on reviewing, assessing, and swiftly taking down content and the measure on internal content policies, for which we have assessed the rights impacts at in the respective sections entitled 'Rights impact' for these measures.

### Freedom of expression and freedom of association

2.240 We note the risk that setting speed-based performance targets can lead to a focus on speed rather than accuracy and note feedback from stakeholders that this could result in incorrect content moderation decisions and the over-removal of content.[167] [168] Such an outcome could have impacts on users' rights to freedom of expression.

2.241 As explained in paragraph 2.233, this risk is mitigated by the flexibility of the measure, which recommends that time targets are set at a level which strikes a balance with accuracy. We consider that this reduces the risks to freedom of expression that may arise with more prescriptive time targets for the removal of illegal content.[169] Additionally, our

[165] We assume that this would require around 30 days of software engineering time, based on the cost assumptions set out in Annex 5.

[166] These cost estimates do not change the approach on which we consulted in our November 2023 Consultation, but add further detail to support our position.

[167] Big Brother Watch response to November 2023 Consultation, pp.4-5; Electronic Frontier Foundation response to November 2023 Consultation, p.8; Meta response to November 2023 Consultation, p.22; Mid Size Platform Group response to November 2023 Consultation, p.8; Pinterest response to November 2023 Consultation, p.7; Reddit response to the November 2023 Consultation, p. 26. We note that Pinterest made a similar point in response to the May 2024 Consultation, p.15.

[168] We note it could equally lead to the over-reporting of content in relation to CSEA or over-takedown of accounts suspected to be run for or on behalf of a proscribed organisation.

[169] Big Brother Watch response to November 2023 Consultation, p.5. We note that Big Brother Watch made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.39.

measure includes the recommendation that services also set performance targets for accuracy, which should mean that both speed and accuracy are considered by services, resulting in greater transparency and consistency in content moderation systems. We consider this potentially would have a positive impact on users' rights to freedom of expression. In applying content moderation systems that efficiently and accurately identify and address illegal content online services will be made safer for users. This could positively impact users' rights to freedom of expression and freedom of association as users would be able to engage with communities and content online more safely.

2.242 While we recognise the risk raised by the Electronic Frontier Foundation that incentivising the swift take down of illegal content could lead to an over-removal of content, the measure requires providers to consider the accuracy of decision-making, which acts as a safeguard for freedom of expression and association and incentivises accurate illegal content removal and reporting.[170] [171]

2.243 The flexibility of the measure also means that providers have scope to set different performance targets for different circumstances – for example, where there is nuance involved with content moderation decisions – to ensure that accuracy is balanced appropriately against speed of decision-making.

2.244 We therefore consider that any interference to users' rights to freedom of expression and freedom of association would be mitigated by the flexibility of the measure, which recommends that time targets are set at a level which strikes a balance with accuracy. Taking this, and the benefits to users and affected persons into consideration, we consider that any impact on rights of freedom of expression and association from this measure is proportionate.

### Privacy and data protection

2.245 We consider the privacy and data protection impacts of this measure to be inextricably linked.

2.246 We note the risk that setting speed-based performance targets can lead to a focus on speed rather than accuracy. This could interfere with users' right to privacy since it may lead to the creation of inaccurate personal data. Therefore, we have designed this measure so that services will need to balance the speed of decisions made with the degree of accuracy, which we consider will mitigate the risk of undue interference with users' rights.

2.247 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. This will be particularly important when making decisions about CSEA content, where an incorrect decision could lead to a user being reported. We consider the measure to be compatible with data protection requirements. We do not consider that it would be appropriate for us to duplicate data protection requirements on the face of the measure in the Codes.

2.248 Overall, and taking the benefits to users and affected persons into consideration, we consider that any impact privacy and data protection rights from this measure is proportionate.

[170] Electronic Frontier Foundation response to November 2023 Consultation, p.8.

[171] We note it could equally lead to the over-reporting of content in relation to CSEA or over-takedown of accounts suspected to be run for or on behalf of a proscribed organisation.

### Who this measure applies to

2.249 We consider that there will be significant benefits to users of applying this measure to providers of multi-risk services, as such providers may have large volumes of content to assess and setting performance targets for their content moderation functions and tracking whether they are met will help them to do this more effectively. Providers that implement this measure are more likely to operate efficient content moderation systems, which in turn play an important role in mitigating the risks of harm to users.

2.250 We note the concern raised by the Online Dating and Discovery Association on the ability of very small businesses to comply with this measure. We accept the cost of this measure could have a significant impact on very small businesses. However, we consider the benefits of applying the measure to providers of multi-risk services are sufficiently important to justify this due to the high probability of illegal content circulating on such services and the fundamental role that effective content moderation plays in protecting users from such content.

2.251 The benefits of this measure will be lower in relation to providers of large low-risk services as they are likely to have a lower volume of content to moderate. However, as explained in 'Our approach to developing Codes measures' we consider that applying this and other measures to such providers will still have significant benefits for users as these services have the potential to affect many users and the nature of illegal content can change over time. In particular, due to the size of such services, there may be a risk of potential illegal content getting lost in the content moderation system without a system in place to keep track of all content requiring assessment. We also consider that large service providers are likely to have sufficient resources to implement this measure.

2.252 As set out in paragraph 2.198, some respondents argued that this measure should apply to all providers of single-risk services in addition to providers of large or multi-risk services.[172] As explained in 'Our approach to developing Codes measures', we expect to consult again on this in Spring 2025. We do not consider it appropriate to delay the start date of the regulatory regime to accommodate a further consultation on this point.

2.253 We are therefore recommending this measure for all providers of large U2U services and all providers of multi-risk services.[173]

## Conclusion

2.254 The analysis above shows that the measure we are recommending is likely to provide significant protections to users from harm and that the costs and impacts on rights that will result from it are proportionate given the scale of the benefits of the measure and the foundational importance of effective content moderation to providers' efforts to protect users from harm. Therefore, we have decided to proceed with the measure broadly as proposed in our November 2023 Consultation, with the following changes:

[172] Age Verification Providers Association response to November 2023 Consultation, p.2; VerifyMy response to November 2023 Consultation, p.6; Yoti response to November 2023 Consultation, p.16.

[173] We note that relevant moderation actions will be different for providers for whom it is currently not technically feasible to take down content.

- time targets apply to the time period for taking relevant content moderation action (where relevant)[174];
- providers can set performance targets for illegal content or an illegal content proxy as appropriate for their service; and
- the measure also now says that providers should balance the **need** to take relevant moderation action swiftly with the **importance** of making accurate moderation decisions. These words **replace the term desirability** which we used in the November 2023 Consultation.

2.255 By implementing this measure, we consider that providers will be clearer on the outcomes they are trying to achieve to protect users and will be able to better configure their systems and processes based on such outcomes.

2.256 This measure will be included in our Illegal Content Codes of Practice for U2U services for terrorism, CSEA and other duties. It is referred to within these Codes as ICU C4.

## Measure on a policy for the prioritisation of content for review

2.257 In the November 2023 Consultation, we proposed that providers should set and apply a policy for the prioritisation of content for review, and, when setting this policy, have regard to the virality of content, potential severity of content and the likelihood content is illegal, including whether it has been flagged by a trusted flagger. We proposed that this measure should apply to all providers of large U2U services and all providers of multi-risk U2U services.

2.258 We considered that prioritisation will help providers make high-quality decisions about what content to prioritise for review, resulting in a material reduction in harm to users.

### Summary of stakeholder feedback[175]

2.259 Our analysis of responses identified several areas where stakeholders felt the measure should be amended. These included (but were not limited to):

- the flexibility of the measure,
- concerns about the practical applicability of the measure,
- severity,
- other factors providers should have regard to in this measure,
- virality,
- trusted flaggers, and
- who this measure applies to.

[174] Targets for CSEA and proscribed organisations content would be needed for the minority of providers for whom take down is technically not feasible.

[175] Note this list in not exhaustive, and further responses can be found in Annex 1.

2.260 We outline these stakeholder concerns in more detail below and address additional stakeholder responses in Annex 1.

## The flexibility of the measure

2.261 Several providers argued that they should have the flexibility to design their policies for the prioritisation of content for review in a way that is most appropriate for their service, including deciding which factors to have regard to.[176]

2.262 We address this point in paragraph 2.289 in the section entitled 'How this measure works.'

## Concerns of the practical applicability of the measure

2.263 Mega said that it is impossible to assess content for the factors listed in the measure, and prioritise content for review having regard to such factors, without reviewing the content first.[177]

2.264 Meta raised concerns that the recommendation that content should be prioritised based on illegality may conflict with the prioritisation of other types of content for review, including content related to the child safety duties in the Act.[178]

2.265 We address these points in paragraphs 2.292 to 2.293 in the section entitled 'How this measure works'.

## Severity

2.266 Refuge noted that the assessment of "severity of content" as a factor providers should have regard to in setting a policy for the prioritisation of content for review is ambiguous and asked if it is intended to measure potential harm to the user.[179]

2.267 We address this point in paragraph 2.301 in the section entitled 'How this measure works'.

2.268 Meta and Snap both commented on our recommendation that providers should use their risk assessments and indications of whether content is suspected to be priority illegal content to indicate the severity of content in the measure.[180] Meta argued against using both of these factors to determine severity, because it would be assessing content for violations of their policies (and whether it should be removed globally), rather than for illegality in the UK in the first instance.[181] It also suggested that for the same reasons, it was not practical for it to prioritise content for review based on the likelihood content is illegal.[182] In contrast, Snap suggested that severity may not align with the priority illegal harms and said that something we consider 'severe' or 'priority' may have no prevalence on a service.[183] It agreed that severity should also be informed by providers' risk assessments.[184]

[176] Google response to November 2023 Consultation, p.37; Meta response to November 2023 Consultation, p.24; Spotify response to November 2023 Consultation, p.6. We note that Google (p.37), Meta (p.24) and TikTok (p.5) made a similar point in response to the May 2024 Consultation.

[177] Mega response to November 2023 Illegal Harms Consultation, p.5

[178] Meta response to November 2023 Consultation, p.24.

[179] Refuge response to November 2023 Consultation, p.12.

[180] Meta response to November 2023 Consultation, p.23; Snap response to November 2023 Consultation, p.11.

[181] Meta response to November 2023 Consultation, p.23.

[182] Meta response to November 2023 Consultation, p.24.

[183] Snap response to November 2023 Consultation, p.11.

[184] Snap response to November 2023 Consultation, p.11.

2.269 We address these points in paragraph 2.306 in the section entitled 'How this measure works'.

## Other factors providers should have regard to in this measure

2.270 Some stakeholders suggested additional factors to those listed in the measure. These included:

- harms to children and the estimated age of users/depicted persons;[185]
- chance of death or serious bodily harm;[186]
- domestic abuse and online violence against women and girls;[187]
- privately sent content;[188]
- the amount of times content is being saved, shared, and commented on; and[189]
- signals of hidden coercion among users, including abuse of platform features like downvotes, mentions and replies.[190]

2.271 We address these points in paragraphs 2.314 and 2.315 in the section entitled 'How this measure works'.

## Virality

2.272 Several stakeholders raised concerns about our recommendation that providers should have regard to virality of content when setting policies for the prioritisation of content for review.

2.273 Snap and UK Interactive Entertainment Industry (Ukie) argued that content was not likely to go viral on their services, and noted that whether it is appropriate for providers to consider this as a factor in setting their policies for prioritisation of content for review varies between different service types.[191]

2.274 We address this in paragraph 2.290 in the section entitled 'How this measure works'.

2.275 Refuge and the Alliance for Countering Crime Online argued against content being prioritised based on virality above other harm-based factors.[192] In its argument, Refuge gave the example of so-called 'honour-based' abuse, which does not need to go viral to risk causing users harm, and would not be prioritised for review based on virality.[193]

[185] Alliance to Counter Crime Online response to November 2023 Illegal Harms Consultation, p.4; Refuge response to November 2023 Consultation, p.12.

[186] Alliance to Counter Crime Online response to November 2023 Consultation, p.4

[187] Refuge response to November 2023 Consultation, p.12.

[188] International Justice Mission's Center to End Online Sexual Exploitation of Children response to November 2023 Illegal Harms Consultation, p.13.

[189] Molly Rose Foundation response to November 2023 Consultation, p.36.

[190] UCL Gender and Tech response to November 2023 Illegal Harms Consultation, p.8.

[191] Ukie response to November 2023 Consultation, p.16; Snap response to November 2023 Consultation, p.11. We note that Ukie made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.40.

[192] Alliance to Counter Crime Online response to the November 2023 Consultation, p.4; Refuge response to November 2023 Consultation, p.12. We note that the Big Brother Watch (p.40) and TikTok (p.5) made similar points in response to the May 2024 Consultation on Protecting Children from Harms Online.

[193] Refuge response to November 2023 Consultation, p.12.

2.276 We address these points in paragraph 2.320 in the section entitled 'Benefits and effectiveness'.

## Trusted flaggers

2.277 Big Brother Watch interpreted trusted flaggers to be referring to the list of trusted flaggers we recommended that some providers should use if they are at high risk of fraud (outlined in chapter 6 of this Volume: 'Reporting and complaints'). [194]

2.278 Meta raised concerns about its practical ability to prioritise reports flagged by trusted flaggers against other types of reports, as, on its service, reports from trusted flaggers are processed in separate dedicated channels that do not operate in tandem with general content moderation systems.[195]

2.279 We address these points in paragraph 2.312 and 2.314 in the section entitled 'How this measure works'.

2.280 Several stakeholders questioned the implication in our measure that content flagged by a trusted flagger should be prioritised over other content. Some stakeholders questioned our assumptions about the accuracy of trusted flagger reports. Big Brother Watch questioned our claim in the November 2023 Consultation that reports from trusted flaggers are likely to be accurate and should therefore be used as an indicator of whether content is likely to be illegal.[196] BILETA suggested that the trusted flaggers which providers use should be Ofcom-approved.[197] An individual requested more information about the evaluation and accountability mechanisms for trusted flaggers.[198]

2.281 Although it said that reports from trusted flaggers were high quality, Snap suggested that such reports do not indicate that harm is widespread and should therefore be prioritised for review.[199]

2.282 We address these points in paragraph 2.323 to 2.325 in the section entitled 'Benefits and effectiveness'.

## Who this measure applies to

2.283 Several stakeholders called for the measure to apply to a wider range of services. The Canadian Centre for Child Protection (C3P) argued that the measure should apply to all U2U services.[200] Age Verification Providers Association, VerifyMy, and Yoti argued that the measure should apply to single-risk services (as well as multi-risk and large services).[201]

2.284 We address these points in paragraph 2.336 in the section entitled 'Who this measure applies to'.

[194] Big Brother Watch response to November 2023 Consultation, p.4.

[195] Meta response to November 2023 Consultation, p.24.

[196] Big Brother Watch response to November 2023 Consultation, p.4. We note that Big Brother Watch made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.40-41.

[197] BILETA response to November 2023 Consultation, p.6.

[198] Are, C. response to November 2023 Consultation, p.7.

[199] Snap response to November 2023 Consultation, p.11.

[200] C3P response to November 2023 Illegal Harms Consultation, p.15. We note that the C3P (p.15). Children's Commissioner for England (pp.59-60); UKSIC (p.36) and Jamie Dean (p.14) made a similar point in response to an equivalent measure proposed in the May 2024 Consultation on Protecting Children from Harms Online.

[201] Age Verification Providers Association response to November 2023 Consultation, p.2; VerifyMy response to November 2023 Consultation, p.6; Yoti response to November 2023 Consultation, p.16.

## Our decision

2.285 We have decided to proceed with the measure broadly as we proposed in the November 2023 Consultation. We have made some small amendments to some of the components of the measure discussed in this section:

- We have clarified that by severity, we mean the severity of harm to UK users if they encounter illegal content on the service.
- We have explicitly stated in the measure that potential harm to children is an aspect of the severity of potential harm caused by content.
- We have also removed the term "virality" from the measure, as we understand that this term is considered to have different meanings by different service providers. Instead, we have said that providers should "have regard to the desirability of minimising the number of United Kingdom users encountering a particular item of illegal content."

2.286 The full text of the measure can be found in our U2U Illegal Content Codes of Practice and is referred to as ICU C5. This measure will be included in our Codes of Practice on terrorism, CSEA and other duties.

## Our reasoning

### How this measure works

2.287 We recommend that providers prepare and apply a policy for the prioritisation of content for review.

2.288 In setting up this prioritisation policy, our measure states that providers should have regard to the following:

- The desirability of minimising the number of UK users encountering a particular item of illegal content;
- The severity of potential harm to UK users if they encounter illegal content on the service, including whether the content is suspected to be priority illegal content, the risk assessment of the service, and the potential harm to children; and
- The likelihood that content is illegal content, including whether it has been reported by a trusted flagger.

2.289 We recognise the concerns expressed by a number of respondents that certain elements of these factors may not be relevant to their service or that prioritisation by reference to certain set of criteria may not be appropriate.[202] For the reasons set out more fully below, we consider that providers should have regard to the factors listed in the measure when *setting* a prioritisation policy. However, the measure is not prescriptive as to how providers should have regard to these factors, and it does not mandate a fixed prioritisation process based on these criteria.

2.290 While we consider providers having appropriate regard to these factors will have important benefits for user safety, we also consider they should have the flexibility to determine

[202] Google response to November 2023 Consultation, p.36; Meta response to November 2023 Consultation, p.24; Spotify response to November 2023 Consultation, p.6. We note that Google (p.6), Meta (p.22) and TikTok (p.5) made a similar point in response to the May 2024 Consultation.

whether, in the light of their risk assessment and the nature of their service, some of the factors are less relevant or not relevant at all. While the measure seeks to secure that providers turn their minds to the factors in deciding how to prioritise content, they retain flexibility in how they do so. We consider that this addresses points made by providers about factors listed in this measure being more or less appropriate for different types of services.[203]

2.291 Many service providers already use systems and processes to help them prioritise content for review. Providers dealing with content moderation on a large scale do not typically review content in chronological order but consider a range of factors, including number of users who are encountering, or have the potential to encounter, content, its severity, and how they became aware of it (for example, as a consequence of a user report or other complaint).[204] Our 'Content Moderation in User-to-User Online Services' report found that Facebook and YouTube both prioritise content that is expected to attract significant viewing.[205] Facebook also gives higher priority to content if the algorithm is confident that moderators will agree it violates content rules. It also prioritises content based on the "severity" or "egregiousness" (and therefore in most cases, the harmfulness) of a suspected violation.[206] Some providers design their complaints forms in ways which ask complainants to categorise the complaint by topic.[207]

2.292 Given this evidence, we consider it would be practical for a provider to prioritise content for review based on the factors listed in our measure without fully reviewing content, in contrast to the argument made by Mega.[208] As we've outlined in paragraph 2.290, it may not be appropriate for all providers to incorporate all the factors listed in the measure into their prioritisation policies for a variety of reasons. However, providers should still consider doing so when designing their prioritisation policies.

2.293 Policies for prioritisation should be applied to all content that is flagged for review, not only suspected illegal content. We are aware that providers already sometimes apply a single prioritisation policy to all content that comes to them for review, including both suspected illegal content and all other types of content suspected to be harmful (including legal content suspected to be harmful to children as mentioned by Meta.[209] It is for providers to determine how to prioritise these different types of content.

[203] Snap response to the November 2023 Consultation, p.11; Ukie response to November 2023 Consultation, p.16. We note that Ukie made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.40.

[204] Ofcom, 2019. Use of AI in Content Moderation. [accessed 25 November 2024]; Meta, 2020. How We Review Content. [accessed 25 November 2024] Meta, 2022. How Meta Prioritises Content for Review. [accessed25 November 2024]. TikTok says it recently started refining their approach to better prioritise accuracy, minimise views of violative content, and remove egregious content quickly. TikTok says it has upgraded the systems that route content for review, to better incorporate a video's expected reach (based on an account's following) when determining whether to remove it, escalate for human review, or take a different course of action. TikTok, 2023. Evolving our approach to content enforcement. [accessed 25 November 2024].

[205] Ofcom, 2023. Content moderation in user-to-user online services: An overview of processes and challenges, p.7. [accessed 25 November 2024].

[206] Ofcom, 2023. Content moderation in user-to-user online services: An overview of processes and challenges, p.20. [accessed 25 November 2024].

[207] Twitch prioritise user reports based on the classification of the report and the severity of the reported behaviour. Twitch, 2023. H1 2023 NetzDG Transparency Report. [accessed 25 November 2024].

[208] Mega response to November 2023 Consultation, p.5.

[209] Meta response to November 2023 Consultation, p.24.

2.294 As outlined in paragraph 2.288, we recommend that providers have regard to certain factors when setting their policies for the prioritisation of content for review. A provider could have regard to these factors for UK users, or any group of users which at a minimum includes UK users. A provider operating globally may choose to set a policy for prioritisation of content for review which applied to all its users, or a larger set of users than just its UK users.

2.295 We explain each of these factors in more detail in the following sections.

### The desirability of minimising the number of UK users encountering a particular item of illegal content

2.296 In setting their prioritisation policies, our measure recommends that providers should have regard to the desirability of minimising the number of UK users encountering a particular item of illegal content.

2.297 A provider could do this by having regard to the degree to which online content has the potential to reach a large number of UK users, or by having regard to the number of UK users who have already encountered content (indicated, for example, by the number of shares, likes or views on a piece of content). A provider has flexibility to have regard to this factor as most appropriate for their service, so long as it considers the desirability of minimising the number of UK users encountering a particular item of illegal content.[210]

2.298 We have amended this factor to remove the term "virality". This is because we understand that this term has different meanings to providers of different services, and we consider that removing it improves clarity.

2.299 We have also added that providers should "**have regard to** the desirability of minimising the number of UK users encountering a particular item of illegal content." Our measure as drafted in the November 2023 Consultation could be interpreted to be recommending that providers make every prioritisation decision based on this factor. This was not the policy intent of this measure. We have therefore amended the measure to clarify that we recommend providers have regard to this factor in setting up their prioritisation policies, alongside the other factors listed in the measure, rather than necessarily making every prioritisation decision based on this.

### Severity of potential harm to UK users

2.300 Providers should have regard to the severity of potential harm to UK users if they encounter illegal content when setting their prioritisation policies. This includes whether content is suspected to be priority illegal content, the risk assessment of the service and the potential harm to children.

2.301 We note Refuge's point that the definition of severity in the measure we proposed in the November 2023 Consultation may not be clear, and we have amended the definition to clarify that we mean the severity of potential harm to UK users if they encounter illegal content.[211]

[210] For example, we know that several providers of larger services consider 'virality' of content when prioritising content for review, including both the 'likely' virality and the 'actual' virality. Meta, 2020. How We Review Content. [accessed 25 November 2024]; Ofcom, 2023. Content moderation in user to-user online services: An overview of processes and challenges. [accessed 25 November 2024]; TikTok, 2023. Evolving our approach to content enforcement. [accessed 25 November 2024].

[211] Refuge response to November 2023 Consultation, p.12.

2.302 We recommend providers should have regard to the severity of potential harm when designing their prioritisation policy, but do not consider they must necessarily have regard to it in every case. We recognise that providers will differ in what they will be able to understand about content before they have looked at it. For example, some providers ask complainants to categorise their complaints. Such providers would be able to write a policy which, for example, prioritised a complaint about CSEA over a complaint about copyright infringement. Some providers may also be able to correlate signals surrounding a complaint or a piece of content with a likelihood of severe harm. It would be appropriate for those providers to apply this knowledge in their prioritisation processes. If a provider chose to use automated content detection to identify content for review, and the automated tools it used could detect aspects of content which are associated with very severe harm, it would also be appropriate to take this into account in its prioritisation processes.

2.303 We know that several service providers already consider the severity of harm when prioritising content for review.[212] There may be degrees of severity that need to be considered within certain kinds of illegal harms.

2.304 We recommend that providers consider whether the content is suspected to be priority illegal content as an indicator of severity, because 'severity' is one of the three factors the UK Government used to determine its list of priority illegal offences.[213]

2.305 We recommend that providers should consider findings from their risk assessments regarding severity of harm when setting content prioritisation policies. Providers which are aware of a particular illegal harm occurring at scale on their service may need to prioritise it for a time until users have learned that the conduct will not be permitted.

2.306 We accept that providers may choose to moderate content based on whether it breaches their terms of service (as outlined in the measures on reviewing, assessing, and taking down content swiftly). However, this violative content may include content that is both legal and illegal. We recommend that providers should have regard to severity of harm to UK users when prioritising content for review, including specifically whether the content is suspected to be priority illegal content, which poses a greater risk of harm. Providers may include other aspects of severity of harm to UK users in their prioritisation decisions (in addition to the factors listed in paragraph 2.288) based on what is appropriate for their service. For example, we do not consider that providers should interpret our explanation of severity to mean that priority illegal content should always be prioritised above the categories of content harmful to children as defined in the Act.

2.307 We agree with the Alliance to Countering Crime Online's suggestion that providers should consider "harms to children" when setting policies for the prioritisation of content for review.[214] We consider this to already be encompassed within our definition of "severity", as we consider the potential for content to be harmful to children to be an important indicator of the severity of harm arising from content given that providers of U2U services are obliged to provide a higher standard of protection for children than for adults within the

[212] Ofcom, 2023. Content moderation in user-to-user online services: An overview of processes and challenges. [accessed25 November 2024].

[213] Department for Digital, Culture, Media & Sport, Home Office, The Rt Hon Nadine Dorries MP, and The Rt Hon Priti Patel MP, 2022. Online safety law to be strengthened to stamp out illegal content. [accessed 25 November 2024].

[214] Alliance to Counter Crime Online response to November 2023 Consultation, p.4.

online safety objectives listed in the Act.[215] We have therefore amended this measure to clarify that potential harm to children is an aspect of the severity of harm to which providers should have regard when setting their content moderation prioritisation policy.

2.308 However, we are not taking forward Refuge's suggestion to include "estimated age of the user/depicted person" as a factor for prioritisation.[216] This would require providers to use technology to estimate the age of people depicted in content. We consider that, to the extent necessary to ensure that providers have regard to technology they currently deploy in assessing potential harm to children, our existing recommendations achieve this.

### The likelihood content is illegal, including whether it has been reported by a trusted flagger

2.309 We recommend that providers have regard to the likelihood that content is illegal, including whether it has been reported by a trusted flagger, in setting their prioritisation policies.

2.310 Providers can be given reasons to suspect that content is illegal in a number of different ways (for example, users may complain about it). Such reports are a valuable way for service providers to find out about illegal content, particularly for those not making extensive use of proactive detection methodologies.

2.311 However, we recognise that users are not always correct when identifying breaches of service providers' content policies.[217] Therefore, we also consider that another indicator of the likelihood that content is illegal is whether it has been reported by a 'trusted flagger'. Trusted flaggers are individuals, non-governmental organisations, government agencies, and other entities that have demonstrated accuracy and reliability in reporting content that violates a provider's terms of service.[218] As a result, they often receive special reporting tools such as the ability to bulk flag content.

2.312 Our Dedicated Reporting Channel measure (in chapter 6 of this Volume: 'Reporting and complaints') sets out instances in which we recommend that providers make a reporting mechanism available to named trusted flaggers in relation to fraud. We note that one stakeholder interpreted this measure as recommending that providers should only consider content flagged by this list of trusted flaggers when considering whether to prioritise content for review.[219] This is not the policy intent in this measure. We have noted stakeholder feedback on our measure for dedicated reporting channels and have clarified in chapter 6 of this Volume: 'Reporting and complaints' that this measure does not prevent the use of the reporting channel for the reporting of other illegal content or intelligence by other trusted flaggers who are assessed by the provider to be sufficiently expert.

2.313 In setting up its prioritisation policy, a provider should have regard to the likelihood that content is illegal, and one factor in determining whether it is illegal will be that it has been reported by a trusted flagger. We have only recommended that providers establish trusted flagger arrangements with entities which we consider can be expected to flag content

[215] See Schedule 4 paragraph 4(a)(vi) and, more generally, section 1(3)(b)(i) of the Act.

[216] Refuge response to November 2023 Consultation, p.12.

[217] For example, Trustpilot's 2021 transparency report says that only 12.4% of consumer user reports in 2021 were deemed to be accurate. Trustpilot, 2021. Trustpilot Transparency Report. [accessed 25 November 2024] Reddit's 2021 transparency report showed that there were 31.3m user reports and it acted on 6.27% of these; the rest were duplicate reports, already actioned, or for content which did not violate its rules. Reddit, 2021. Transparency Report 2021. [accessed 25 November 2024].

[218] European Commission, 2017. Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms. [accessed 24 November 2024].

[219] Big Brother Watch response to the November 2023 Consultation, p.4.

correctly, and we do not recommend that providers establish relationships with trusted flaggers unless they are asked to do so by the trusted flagger concerned. Our measure also leaves it open to providers to have a policy for prioritising content for review which is not based on this factor, so long as, in setting their policy, they have considered whether and how to prioritise flags from trusted flaggers.

2.314 A provider could set up its policy for the prioritisation of content flagged by trusted flaggers in a variety of ways, including having a separate team to review this content than the team it uses to review other content flagged in (for example) user reports.

### Other factors

2.315 We have not decided to add any of the additional factors stakeholders suggested providers should have regard to in setting their prioritisation policies. Recommending that providers have regard to a particular harm (and therefore implying this harm to be of greater importance than others) in setting policies for the prioritisation of content could give rise to a significant risk of unintended consequences. For example, this could lead to other harms that may be more prevalent on the service being deprioritised, remaining on the service for longer and causing greater harm to users. For the reasons outlined in paragraph 2.307, we consider an exception to this to be potential harms to children as an aspect of the potential severity of content. This is due to the explicit duties on providers to prioritise this type of harm as set out in the Act.[220]

2.316 However, this measure gives providers the flexibility to incorporate our recommended prioritisation factors into their own prioritisation frameworks as appropriate to their services. Service providers have the flexibility to consider other factors for prioritisation as appropriate for their service, and to make different types of prioritisation decisions for different pieces of content and different types of services.

## Benefits and effectiveness

2.317 Prioritisation decisions can have a material impact on the amount of harm a piece of content causes to users. For example, if a provider chose to review a series of relatively minor pieces of illegal content which were not viewed by many (or any) people before it reviewed a piece of extremely harmful illegal content that was viewed by a large number of people, this decision could result in significant harm to users.

2.318 We consider that setting a policy for the prioritisation of content for review that considers the factors outlined in paragraph 2.288 will result in providers making better quality decisions about what content to prioritise for review to protect users, as opposed to reviewing complaints in a chronological order.

2.319 If illegal content is reaching a higher number of UK users than is typical within a given timeframe, or has the potential to reach a high number of users, then it has the potential to cause harm to larger audiences. We therefore consider that service providers for whom this factor is relevant will achieve better outcomes for users if they have regard to the number of users encountering, or having the potential to encounter, illegal content when setting up their prioritisation policies.

2.320 We note stakeholder arguments that it is important to balance this alongside other factors, as setting a policy for prioritisation based on this factor alone may mean other harms are

[220] See Schedule 4 paragraph 4(a)(vi) and, more generally, section 1(3)(b)(i) of the Act.

missed.[221] For example, CSAM is not typically encountered by a large number of users but the harm caused by such content is severe. Similarly, content constituting harassment and threats or intimate image abuse may be targeted at an individual and may not be encountered by a large number of users but can cause severe harm to the individual concerned (and can be particularly harmful to women and girls). We consider our recommendation that providers have regard to the other factors listed in our measure alongside the factor relating to minimising the number of users encountering a particular item of illegal content, particularly the factor on the severity of potential harm to UK users if they encounter illegal content on the service, will mitigate this risk.

2.321 When setting policies for prioritising content for review, providers should have regard to the severity of the potential harm to UK users if they encounter illegal content on the service. This relates to content suspected to be priority illegal content and content that is potentially harmful to children, each of which the provider will assess in its risk assessments.[222] While we recognise that some providers will prefer to consider content by reference to their own terms of service, the underlying harms with which we are concerned are the harms regulated by the Act. Therefore, in order to prepare a prioritisation policy, a provider will need to consider those harms. For illegal harms, if a provider wishes to make individual prioritisation decisions about specific items of content by reference to its own terms of service, it first needs to have considered (as a part of preparing its policy on prioritisation) how well those terms of service correspond to the severity of the harm from illegal content.[223]

2.322 The likelihood that content is illegal content is highly relevant to whether further review is needed and how quickly it should take place. The fact that a complaint comes from a trusted flagger or another expert body is also of relevance in determining what priority to give it, as such complaints are likely accurate and reflective of the trusted flagger's assessment of harm.

2.323 We recognise some stakeholders' concerns that there could be cases in which trusted flaggers reports are not accurate or need extra checks in place to ensure their accuracy.[224] In relation to the trusted flaggers we have specifically recommended, we consider that we have chosen entities that can be expected to make accurate reports, as they are public entities with an expertise and competence in relation to tackling fraud.[225]

2.324 More generally, we do not expect a provider to prioritise the reports of trusted flaggers in relation to which they do not have confidence the reports will be accurate (although we

[221] Alliance to Counter Crime Online response to the November 2023 Consultation, p.4; Refuge response to November 2023 Consultation, p.12; We note that Big Brother Watch (p.40) and TikTok (p.5) made similar points in response to the May 2024 Consultation on Protecting Children from Harms Online.

[222] We recognise that the illegal harms safety duty will come into force before the protection of children safety duty. We do not expect providers to inappropriately prioritise illegal content over content harmful to children in the period before the children's safety duty comes into force.

[223] Meta response to November 2023 Consultation, p.23; Snap response to November 2023 Consultation, p.11.

[224] Are, C. response to November 2023 Consultation, p.7; BILETA response to November 2023 Consultation, p.6; Big Brother Watch response to November 2023 Consultation, p.4. We note that Big Brother Watch (pp.40-41) made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, pp.40-41.

[225] Our recommendation in relation to the trusted flaggers we specify includes a recommendation that providers publish processes relating to the establishment of a dedicated reporting channel for trusted flaggers, covering any relevant procedural matters. It is open to a provider which has concerns about accuracy of flags to provider for escalation and dispute resolution in its processes.

would expect a provider to carefully consider why, in that case, the entity was considered a trusted flagger). It is for providers to factor in the reliability of their trusted flaggers when they consider the weight to put on their complaints in their prioritisation processes. By definition, we expect trusted flaggers are generally more likely to be reliable than reports from people who are not trusted flaggers, but that does not mean they are always correct.

2.325 We consider Snap's argument that trusted flagger reports do not always indicate that content is widespread on the platform, and should therefore be prioritised, is already addressed in the measure.[226] We are not recommending that service providers necessarily consider trusted flaggers for every prioritisation decision they make, but instead recommend them as one way for service providers to establish if content is likely to be illegal when having regard to this factor in setting their policies for the prioritisation of content for review. As stated in paragraph 2.288, we also recommend service providers have regard to the desirability of minimising the number of users encountering a particular item of illegal content and the severity of potential harm caused by content, as well as the likelihood content is illegal in this measure. This measure also gives providers the flexibility to have regard to any other factors they deem appropriate for their service, in setting these prioritisation policies.

2.326 Ultimately, the amount of harm caused by a particular piece of illegal content is likely to be in significant measure a function of the severity of the content and the number of people exposed to it. This being the case, benefits to users are likely to be maximised where providers have regard to the factors set out above when prioritising potentially harmful content for review.

### Costs and risks

2.327 The creation of a prioritisation policy will not in itself have an impact on the overall amount of content that providers need to review. However, there will be costs incurred in designing and applying the prioritisation policy. The largely one-off costs of designing the prioritisation policy may take several weeks of full-time work and involve legal, regulatory, and ICT staff, as well as experts in harms and online safety. Agreeing new policies may also require input from senior management, which would add to the upfront costs. Applying the prioritisation policy may require system changes (such as ensuring the potential severity of content is taken into account when content is reviewed by content moderators or ensuring that content from trusted flaggers is suitably prioritised). There may be material one-off costs in making these changes. Service providers will also incur one-off costs in designing and setting up suitable performance metrics and targets.

2.328 There are likely to be some smaller ongoing costs in ensuring that the prioritisation policy is still reflected in system design and reviewing it when appropriate. We consider that all these costs will be mitigated by the measure not specifying exactly how service providers should prioritise content, giving them some flexibility how they achieve this.

2.329 Since our November 2023 Consultation, we have further analysed these costs for the purposes of our May 2024 Consultation. We anticipate that designing and setting up a relatively simple prioritisation framework (for example, to suit a smaller service that has identified itself as being at medium or high risk for two types of illegal harm, and which has

[226] Snap response to the November 2023 Consultation, p.11.

a limited quantity of content to review) could cost approximately £4,000 to £7,000.[227] While this cost estimate relates to developing a prioritisation system relating to content harmful to children, we expect that the costs of developing such a system relating to illegal harms could be similar for many smaller providers because the development process, staff involvement, and time required is likely to be similar.

2.330 For a larger and more complex service which has identified itself as being high-risk for many types of illegal harms (and is likely to have many different metrics that can indicate virality, severity, and suspected type of content), costs could be substantially higher, potentially reaching tens of thousands or more.[228] As the total amount of content reviewed will not change due to this measure, it is not clear whether establishing a framework for prioritising content would impose material ongoing content moderation costs on service providers (compared to simply reviewing complaints in chronological order). For service providers that do not already do this, having a clear prioritisation framework may help them deploy their resources more effectively.

## Rights impact

2.331 This proposed measure should be seen as part of a package of measures relating to content moderation for illegal content, including the measures on reviewing, assessing and swiftly taking down content and the measure on internal content policies, for which we have assessed the rights impacts in the respective sections entitled 'Rights impact' for these measures.

### Freedom of expression and freedom of association

2.332 We do not consider that setting and applying a content prioritisation policy would, in itself, have any specific adverse impacts on users' or services' rights to freedom of expression or association. It may have a positive impact on the right to freedom of expression, as the recommendations of the measure mean that harm would be a factor in service providers' decision making and that users would be able to engage with communities and content online more safely. Overall, and taking the benefits to users and affected persons into consideration, we consider that any impact on rights of freedom of expression and association from this measure is proportionate.

### Privacy and data protection

2.333 We consider that setting and applying a prioritisation policy would only have additional impacts on users' privacy or personal data rights beyond those already considered, to the extent that it involved a further use of private information or processing of personal data by the provider concerned. However, any such extra processing would need to be carried out in compliance with applicable privacy and data protection laws. Taking this, and the benefits to users and affected persons into consideration, we consider that any impact on privacy and data protection rights from this measure is proportionate.

## Who this measure applies to

2.334 We consider that the benefits of adopting a prioritisation framework for providers of multi-risk services make it proportionate for these providers to incur the costs of doing so. These

[227] Assuming this would require three weeks FTE from professional occupations (legal, regulatory, ICT) and one day from senior management, based on our salary assumptions as set out in Annex 5.

[228] These cost estimates do not change the approach on which we consulted in our November 2023 Consultation but add further detail to support our position.

providers are likely to have a large quantity of potentially illegal content to review, and significant potential harms may arise from this content. Having a prioritisation framework in place will help service providers focus their content moderation resources on reviewing pieces of content that are more likely to cause severe harm and affect many users.

2.335 The benefits of this measure will be lower in relation to providers of large low-risk services as the scope to reduce harm will be more limited. However, as explained in 'Our approach to developing Codes measures' we consider that applying this and other measures to such services will have benefits for users as these services have the potential to affect many users and the nature of illegal content can change over time. In particular, we note that providers of large services may still have substantial volumes of content to moderate. Having a prioritisation framework in place will help ensure that any risk of illegal content on such services can be dealt with quickly, reducing the resulting harms. Providers of large services are also likely to have sufficient resources to develop or adjust policies in line with the measure. We therefore consider that it is proportionate to apply this measure to providers of large, low-risk services.

2.336 As set out in paragraph 2.283, some stakeholders argued that this measure should also apply to providers of single-risk services, or to all providers (regardless of risk level).[229] [230] We consider that the benefits of having a prioritisation framework are likely to be materially lower for smaller services that are low-risk for all kinds of illegal harm. Because the volume of potentially illegal material such services will need to review will be very materially lower, they are much less likely to face difficult prioritisation decisions. We therefore maintain that it would not be proportionate to extend this measure to such service providers.

2.337 We note the arguments that this measure should apply to some or all providers of single-risk services. As explained in 'Our approach to developing Codes measures', we expect to consult again on this in Spring 2025. We do not consider it appropriate to delay the start date of the regulatory regime to accommodate a further consultation on this point.

2.338 We are therefore recommending this measure for all providers of large U2U services and all providers of multi-risk U2U services.

## Conclusion

2.339 Our analysis shows that the measure we are recommending is likely to provide significant protections to users from harm, the costs are proportionate and it will have no additional negative impacts on rights. Therefore, we have decided to proceed with the measure largely unchanged from the measure we proposed in our November 2023 Consultation, except for some small changes:

- We have clarified that by severity we mean the severity of harm to United Kingdom users if they encounter illegal content on the service.

[229] C3P response to the November 2023 Consultation, p.15. We note that the C3P (pp.21-22); Children's Commissioner for England (pp.59-60), UKSIC (p.36), and Jamie Dean (p.14) made a similar point in response to an equivalent measure proposed in the May 2024 Consultation.

[230] Age Verification Providers Association response to November 2023 Consultation, p.2; VerifyMy response to November 2023 Consultation, p.6; Yoti response to November 2023 Consultation, p.16.

- We have explicitly stated in the measure that potential harm to children is an aspect of the severity of potential harm caused by content.
- We have also removed the term "virality" from the measure, as we understand that this term is considered to have different meanings by different service providers. Instead, we have said that providers should have regard to the desirability of minimising the number of United Kingdom users encountering a particular item of illegal content.

2.340 Providers of large and multi-risk services should prepare and apply a policy for the prioritisation of content for review. In setting up this policy, a provider should have regard to the desirability of minimising the circumstances in which the number of users encountering a particular piece of illegal content can increase exponentially over time; the severity of potential harm to UK users if they encounter illegal content on the service and whether content is likely to be illegal, including whether it is reported by a trusted flagger. By implementing this measure, we consider that providers will make higher quality decisions about what content to prioritise for review to protect users.

2.341 This measure will be included in our Illegal Content Codes of Practice on terrorism, CSEA and other duties. It is referred to within these Codes as ICU C5.

## Measure on resourcing content moderation

2.342 In the November 2023 Consultation, we proposed that service providers should ensure their content moderation teams are resourced to give effect to the measures on internal content policies and performance targets. In doing this, we proposed that providers should have regard to at least (1) the propensity for external events to lead to a significant increase in demand for content moderation, and (2) the particular needs of their UK user base as identified in their risk assessments, in relation to languages. We proposed that this measure should apply to all providers of large U2U services and all providers of multi-risk U2U services.

2.343 We explained our view that sufficiently resourcing content moderation to achieve their performance targets will help providers review potentially illegal content faster and more accurately. The effectiveness of our recommendation on performance targets also relies on providers resourcing their content moderation functions to meet those targets.

### Summary of stakeholder feedback[231]

2.344 In addition to those stakeholders who expressed broader support for the full package of content moderation measures outlined in paragraph 2.14, several stakeholders expressed support specifically for this measure.[232]

2.345 There were some other themes in responses including (but not limited to):

- external events;
- language;
- outsourced content moderation; and

[231] Note this list is not exhaustive, and further responses can be found in Annex 1.

[232] Mencap response to November 2023 Illegal Harms Consultation, p.8; Meta response to November 2023 Illegal Harms Consultation, p.24; NSPCC response to November 2023 Consultation, p.21; Open Rights Group response to November 2023 Illegal Harms Consultation, p.2.

- who the measure applies to.

2.346 We outline these stakeholder views in more detail below, and address additional stakeholder responses in Annex 1.

## External events

2.347 While Snap and Meta both agreed that it was possible for providers to have strategies in place to respond to external events, they emphasised that some events were unpredictable and may require more tailored, ad hoc responses.[233] The OSA Network raised concerns that our reference to external events in the measure did not include unexpected events like terrorism.[234]

2.348 We address these points in paragraphs 2.372 and 2.373 in the section entitled 'Benefits and effectiveness'.

## Language

2.349 Snap, the NSPCC, and Centre for Competition Policy supported our recommendation that in resourcing their content moderation functions, providers should have regard to the particular needs of their UK user bases as identified in their risk assessments in relation to languages.[235] The NSPCC cited evidence of how limited numbers of staff who understand local language and cultural references has resulted in challenges in tackling disinformation on X (formerly Twitter).[236]

2.350 Meta agreed that providers should have regard to the language needs of their UK user bases in resourcing their content moderation function, while also highlighting the complexity of this process. It shared that it uses content moderation teams that provide global coverage, whose resource can be redeployed to different countries during surges in need for moderation in particular languages.[237]

2.351 While not in direct reference to this measure, some stakeholders also made more general comments about the need for people working in moderation to have an understanding of the language of content they moderate.[238]

2.352 Some stakeholders argued that the recommendation for resourcing content moderation for different languages should go further. BILETA suggested that moderation staff should not only be able to speak relevant languages fluently but be familiar with "prejudicial terminology".[239] In response to the equivalent measure in the May 2024 Consultation, the

[233] Meta response to November 2023 Consultation, p.25; Snap response to November 2023 Consultation, p.11. We note that Meta made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.23.

[234] OSA Network response to November 2023 Consultation, p.97.

[235] Centre for Competition Policy response to November 2023 Consultation, p.17; NSPCC response to November 2023 Consultation, p.21; Snap response to November 2023 Consultation, p.11.

[236] NSPCC response to November 2023 Consultation, p.21.

[237] Meta response to November 2023 Illegal Harms Consultation, p.25. We note that Meta made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.23.

[238] BILETA response to November 2023 Consultation, p.6; Electronic Frontier Foundation response to November 2023 Illegal Harms Consultation, p.10; Open Rights Group response to November 2023 Consultation, p.2. We note that Northern Ireland Commissioner for Children and Young People made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.32.

[239] BILETA response to November 2023 Consultation, p.6.

Conscious Advertising Network argued that we should mandate providers to moderate in all languages spoken in the UK, not just English.[240]

2.353 We address these points in paragraphs 2.374 to 2.379 in the section entitled 'Benefits and effectiveness'.

### Outsourced content moderation

2.354 Refuge argued that outsourced content moderation resource should be included in our recommendations.[241] 5Rights Foundation argued that our cost assumptions for human moderation resource do not take into account the outsourcing of these roles to other countries, in which salaries can be lower.[242]

2.355 We address these points in paragraph 2.384 in the section entitled 'Costs'.

### Who this measure applies to

2.356 VerifyMy argued the measure should apply to providers of services with a single risk (as well as providers of large and multi-risk services).[243]

2.357 Barnardo's suggested that the measure should apply to all services.[244]

2.358 We address these points in paragraph 2.389 and 3.390 in the section entitled 'Who this measure applies to'.

## Our decision

2.359 We have decided to proceed with the measure as proposed in our November 2023 Consultation. The full text of the measure can be found in the U2U Illegal Content Codes of Practice and will be included in our Codes of Practice for U2U services on Terrorism, CSEA and other duties. It is referred to as ICU C6.

## Our reasoning

### How this measure works

2.360 We recommend that providers of large and multi-risk services should resource their content moderation functions adequately to give effect to their internal content policies and performance targets.[245] We recommend that in doing so, they should have regard to at least (i) the propensity for external events to lead to a significant increase in demand for content moderation on the service, and (ii) Resourcing for the particular language needs of UK user bases.

[240] Conscious Advertising Network response to May 2024 Consultation on Protecting Children from Harms Online, p.9.

[241] Refuge response to November 2023 Consultation, p.12.

[242] 5Rights Foundation response to November 2023 Consultation, p.20.

[243] VerifyMy response to November 2023 Consultation, p.6.

[244] Barnardo's response to November 2023 Consultation, p.15. We note that the C3P (pp.21-22); Children's Commissioner for England (pp.59-60); Molly Rose Foundation (p.41) and UKSIC (p.36) made a similar point in response to an equivalent Measure proposed in the May 2024 Consultation.

[245] As we recommend in accordance with measures ICU C3 and ICU C4 respectively.

### The propensity for external events to lead to a significant increase in demand for content moderation on the service

2.361 In instances where systems may need to deal with sudden significant increases in illegal content or unexpected harm events, redeploying resources to do so may draw resources away from another part of the system. It is beneficial for providers to consider the potential for sudden significant increases in problematic (and potentially illegal) content when determining how to resource their content moderation functions.

2.362 Information obtained from service providers' risk assessments, tracking evidence of new kinds of illegal content and other relevant sources of information could be used to understand where and when some such occurrences might happen.

### Resourcing for the particular language needs of UK user bases

2.363 The provider should consider the particular language needs of its UK user base as identified in its risk assessment. This means that if a large proportion of the UK userbase is likely to use the service in certain languages, then the content moderation function should be equipped to moderation content in those languages accordingly.

2.364 We expect that providers should be prepared to adapt to changing prevalence in languages across their UK users.

2.365 We outline our reasoning for these recommendations in paragraphs 2.374 to 2.379 in the section entitled 'Benefits and effectiveness'.

## Benefits and effectiveness

2.366 Providers with large volumes or many different types of content to review are unlikely to be able to keep users safe using ad hoc methods (for example, by ensuring that whichever member of senior management is available reviews complaints as they come in). Providers are likely to need specific resources to handle complaints and may need to adjust their overall resources and how they use them based on what is happening on their service.

2.367 We therefore consider that adequate resourcing of content moderation functions will result in providers making more accurate and timely decisions about whether content is illegal content, and, where applicable, whether they should remove that content.[246] We would expect this to result in a material reduction in harm to users and deliver significant benefits.

2.368 This aligns with the evidence discussed in the Register of Risks which concerns how limited resourcing, time pressures, and large or fluctuating volumes of content requiring moderation can contribute to increased risk. It also highlights specific instances where this is reported to have happened – for example, some studies have concluded that the reduction in content moderation capacity at X (formerly Twitter) led to a major increase in the quantity of antisemitic content on the service.[247]

2.369 Setting performance targets in relation to the speed and accuracy of a U2U content moderation function will not protect users unless service providers ensure they have sufficient resources and deploy them effectively to meet their targets. We expect there to

[246] Ofcom research suggests that, all other things being equal, a provider may be able to reduce the 'turnaround time' between content being uploaded, reviewing and actioned by hiring more moderators, thereby reducing the amount of time that potentially harmful or violative content is 'live'. Ofcom, 2023. Content moderation in user-to-user online services, p.26. [accessed 25 November 2024]

[247] CASM Technology and ISD, 2023. Antisemitism on Twitter Before and After Elon Musk's Acquisition. [accessed 25 November 2024].

be significant benefits from service providers resourcing their content moderation functions in a way that allows them to meet their performance targets.

2.370 We also consider that there are factors to which service providers should have regard when deciding how to resource their content moderation function. We explain why each factor is important in paragraphs 2.371 to 2.379.

### The propensity for external events to lead to a significant increase in demand for content moderation on the service

2.371 Evidence suggests that service providers need to build flexibility into their content moderation functions for them to be effective.[248] [249] In response to the 2022 Illegal Harms Call for Evidence, British Sustainable Business Network and Consultancy (BSR) stressed the importance of service providers "investing in the capability to scale-up/scale-down on short notice to respond to crisis events that can result in sudden spikes in illegal content".[250] We consider it will be beneficial for service providers to anticipate potential spikes in demand for content moderation due to external events and adjust their resources accordingly.

2.372 In relation to the OSA Network's comments regarding external events, our intention in specifying "external events" is that this wording is sufficiently broad to encompass both expected events and unexpected events.[251] By way of example only, expected events might include planning around election campaigns or major sporting events (where a provider's risk assessment might highlight the risk of foreign interference or the potential for illegal hateful content). Unexpected events might include contingency planning in the case of terrorist attacks or civil unrest.[252]

2.373 We note the responses from some stakeholders highlighting that some external events, especially unexpected events, may need individualised responses that could be difficult to plan for in advance in terms of resourcing.[253] However, we still consider that there are important benefits to providers having regard to potential spikes in demand driven by all types of external events, including unexpected events, in resourcing their content moderation functions. In instances where systems may need to deal with sudden significant

[248] For example, a report by the Alan Turing Institute that tracked abuse of Premier League football players on Twitter during the 2021–22 Premier League season, found that hate speech peaked following key events. The Alan Turing Institute, 2022. Tracking abuse on Twitter against football players in the 2021 – 22 Premier League Season. [accessed 25 November 2024].

[249] For example, following the start of the 2023 Israel-Gaza war, providers of U2U services and other organisations reported an increase in harmful content online, including that which encourages hate and incites violence and graphic violent videos and images. Amnesty International, 2023. Global: Social media companies must step up crisis response on Israel Palestine as online hate and censorship proliferate. 27 October 2023. [accessed 25 November 2024]. Scott, M., Graphic videos of Hamas attacks spread on X. Politico, 9 October 2023. [accessed 25 November 2024]. Meta Oversight Board, 2023. Hostages Kidnapped from Israel. [accessed 25 November 2024]. Meta Oversight Board, 2023. Al-Shifa Hospital. [accessed 25 November 2024]. We are aware that some providers adjusted their content moderation capabilities in response to an increase in hate content following the start of the 2023 Israel-Gaza war.  TikTok, 2023. Our continued actions to protect the TikTok community during the Israel-Hamas war. [accessed 25 November 2024]. Meta, 2023. Meta's Ongoing Efforts Regarding the Israel-Hamas War. [accessed 25 November 2024].

[250] BSR response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation.

[251] OSA Network response to November 2023 Consultation, p.97.

[252] We outline evidence of how such types of events, both expected and unexpected, manifest on services in our Register of Risks.

[253] Meta response to November 2023 Consultation, p.25 and Snap response to November 2023 Consultation, p.12. We note that Meta made a similar point in their response to the May 2024 Consultation on Protecting Children from Harms Online, p.23.

increases in illegal content or unexpected harm events, redeploying resources may deplete another part of the system. Service providers that have contingency plans in place to ensure that illegal content across the system is dealt with efficiently are more likely to be effective in protecting users from harm in these instances. We recommend that providers have regard to the potential for spikes in demand when determining how to resource their moderation functions.

### Resourcing for the particular language needs of providers' UK user bases

2.374 Given the large number of languages that are spoken in the UK and the fact that some services may be target at specific communities of language speakers, content that has the potential to cause harm to UK users may be posted in multiple languages. Harm is likely to be reduced where service providers ensure their content moderation functions include the language skills needed to review potentially illegal content that could affect these users. In addition to the stakeholders who expressed their support for this part of the measure in response to the November 2023 Consultation, several stakeholders responded to our 2022 Illegal Harms Call for Evidence and our 2023 Protection of Children Call for Evidence stressing the importance of being able to moderate in different languages. They also noted the importance of moderators having a knowledge of cultural context, to enable them to better understand the context relevant for the content being reviewed.[254]

2.375 We are aware that several service providers already consider the language in which content is posted and ensure they have the language expertise within their moderation systems to deal with such content (using both human and automated methods to do so).[255] Facebook and Instagram have global content review teams that review content in more than 70 languages, 24 hours a day, seven days a week.[256] TikTok moderates content in more than 70 languages and provides information about the primary languages in which its moderators work.[257] In its response to the 2022 Illegal Harms Call for Evidence, Glassdoor stated that it uses proprietary technology to analyse all English and non-English language content.[258] Snap told us that it uses content moderation teams worldwide, including both internal teams and third-party vendors to cover languages where the service is available across the

[254] Glassdoor response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation.; BSR response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation.; Glitch response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation.; Global Partners Digital response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation.; Common Sense Media response to 2023 Ofcom Protection of Children Call for Evidence. Glitch response to 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation. In advice to the United Nations Special Rapporteur on Minority Issues, in relation to hate speech specifically, Carnegie UK said, 'companies should ensure that, proportionate to risk they have sufficient moderators trained on language and cultural considerations to combat hate speech.' Carnegie UK, 2021. Ad hoc advice to the United Nations Special Rapporteur on Minority Issues. [accessed 25 November 2024].

[255] "The social media companies said they moderated content or provided fact- checks in many languages: more than 70 languages for TikTok, and more than 60 for Meta, which owns Facebook. YouTube said it had more than 20,000 people reviewing and removing misinformation, including in languages such as Mandarin and Spanish; TikTok had thousands. The companies declined to say how many employees were doing work in languages other than English." Hsu, T. Misinformation Swirls in Non-English Languages Ahead of Midterms. The New York Times, 12 October 2022. [accessed 26 November 2024].

[256] Facebook, 2023. DSA transparency report. [accessed 25 November 2024]. Instagram, 2023. DSA transparency report. [accessed 25 November 2024].

[257] TikTok, 2023. TikTok's DSA Transparency Report. [accessed 25 November 2024].

[258] Glassdoor, 2022. Glassdoor response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation.

world.[259] Twitter (now X) told us it has teams spread around the world specifically trained to provide coverage in the languages it serves on the service.[260] Bumble told us it uses a suite of bespoke artificial intelligence (AI) moderation tools that cover over 100 languages.[261] TikTok told us it uses local experts to help develop and constantly refine its policies. It also has local language experts helping moderation teams.[262]

2.376 We are also aware of the harm caused by providers not adequately accounting for the language of their users in the resourcing of their content moderation systems and processes, as raised by the NSPCC in outlining its support for this recommendation.[263] We note suggestions that many providers do not currently have sufficient language expertise in place to deal with the variety of languages or nuances with languages or cultural references on their services, which can lead to content moderation systems failing to identify illegal or harmful content.[264]

2.377 We consider that our recommendation helps address more general stakeholder arguments about the need for people working in moderation to have an understanding of the language of content they moderate.[265]

2.378 However, we do not consider it would be appropriate at this stage for our recommendation on language resourcing to be more prescriptive. The language expertise required to deal with the risk of harm in a particular language, including understanding of prejudicial terminology, is likely to differ from service to service based on a number of factors, including user base, content type, and functionality. There are likely to be trade-offs, particularly in relation to languages spoken less widely in the UK, as the value of having content moderated by a fluent speaker of the language may or may not be greater than the value of having it moderated by someone with an understanding of the UK and of UK definitions of illegal content. For this reason, we have not made recommendations regarding the exact language expertise and resourcing services should have in place.

2.379 We also consider that our measures on training individuals working in moderation and on providing materials to volunteers will address cases where a lack of contextual understanding beyond language (such as a lack of understanding of prejudicial terminology) creates a gap in the understanding of individuals working in moderation' in relation to a particular illegal harm on a service. We discuss this in more detail when explaining our reasoning for this in paragraph 2.441 in the measure on training individuals working in moderation.

[259] Snap response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.
[260] Twitter response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.
[261] Ofcom/Bumble meeting, 25 October 2022.
[262] Ofcom/TikTok meeting, 7 February 2022.
[263] NSPCC response to November 2023 Consultation, p.21.
[264]AACJ, 2022. Fake News and the Growing Power of Asian American Voters: What this Means for 2022 Midterm Elections. [accessed 25 November 2024]; State of the Internet's Languages, 2022. State of the Internet's Languages Report. [accessed 25 November 2024]; Global Partners Digital, 2022. Marginalised Languages and the Content Moderation Challenge. [accessed25 November 2024]; Oversight Board, 2022. Oversight Board Annual Report 2021. [accessed 25 November 2024]; AI4Dignity, 2021. Artificial Intelligence, Extreme Speech, and the Challenges of Online Content Moderation. [accessed 25 November 2024]
[265] BILETA response to November 2023 Consultation, p.6; Electronic Frontier Foundation response to November 2023 Consultation, p.10; Open Rights Group response to November 2023 Consultation, p.2. We note that Northern Ireland Commissioner for Children and Young People made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.32.

## Costs and risks

2.380 The costs of resourcing content moderation systems in line with this measure are likely to be substantial and ongoing. We expect costs to vary depending on the size of the service, the policies developed by the provider, and the nature and volume of illegal content on the service. In general, we would expect costs to be lower in relation to smaller services and higher in relation to larger services. However, providers of small services may face higher costs if they are high-risk.

2.381 Providers have flexibility over the mix of human and automated content moderation they use, and the type of detection and review processes chosen is likely to influence costs. Automated moderation processes, such as machine learning solutions for artificial intelligence (AI), require both one-off infrastructure investment and time input from various ICT professionals. Ongoing costs may also be incurred from system updates and licensing fees. Providers of large services may be able to develop these in-house, but the costs of doing so can be high. Because of this, providers of smaller services may outsource development to a third party or use off-the-shelf third-party solutions.

2.382 If content moderation involves human moderators, resourcing costs will primarily depend on how many moderators are needed. To be effective, human moderators may require specific tools and/or training to detect and review content (as outlined in the measure on provision of training and materials to individuals working in moderation). They may also need an ICT support team. Service providers may also decide to offer mental health support and other wellbeing benefits to their content moderators and other individuals working on content moderation, which would add to costs.[266] Some providers might also require a separate review process for more complex illegal content cases, which may require legal input.

2.383 Providers that already have policies and processes in place that are sufficient to implement this measure would not need to incur any additional costs unless they withdraw these policies and processes.

2.384 As set out in paragraph 2.354, two stakeholders argued that providers should take into account outsourced content moderation resources.[267] To clarify, we have already considered inclusion of outsourced content moderation in the measure and note that content moderation costs could be lower where this is the case.

## Rights impact

2.385 This should be seen as part of a package of measures relating to content moderation for illegal content, including the measures on reviewing, assessing, and swiftly taking down content and the measure on internal content policies, for which we have assessed the rights impacts in the relevant sections of this chapter. We do not consider that the measure will have any additional negative impact on users' rights. Appropriately resourcing content moderation is likely to have positive impacts on users' rights, and safeguard those rights, because mistakes are less likely and because the result should be that users feel safer using

[266] Where content moderation is performed by employees of a provider, the provider will need to consider their duty of care to these employees and which the provider may consider involves offering such support and benefits.

[267] 5Rights Foundation response to November 2023 Consultation, p.20; Refuge response to November 2023 Consultation, p.12.

the service. Overall, and taking the benefits to users and affected persons into consideration, we consider that any impact on rights from this measure is proportionate.

## Who this measure applies to

2.386 We have decided to apply this measure to providers of large U2U services and providers of multi-risk U2U services. Given the amount of content posted on large U2U services, providers of such services often get large volumes of content flagged to them as potentially illegal (or otherwise harmful). Providers of smaller multi-risk services are also likely to have many different types of content to moderate at once. This means that providers of both types of service are unlikely to be able to keep users safe using ad hoc methods (for example, by ensuring that whichever member of senior management is available reviews complaints as they come in). Providers are likely to need specific resources to handle complaints and may need to adjust their overall resources and how they use them based on what is happening on their service.

2.387 Although our analysis suggests that this measure could impose significant costs on providers, we consider that well-resourced content moderation functions will deliver very significant and important benefits, particularly in relation to multi-risk services. We expect this to result in a material reduction of harm (compared to a counterfactual scenario where the service operates with a lower level of resource that may be insufficient to fully implement its provider's internal content moderation policies and achieve the targets set). Overall, we conclude that the benefits associated with the measure are so significant, and adequately resourcing content moderation teams is so fundamental to user safety, that it is proportionate to apply the measure to the service types in question despite the potentially significant costs.

2.388 As explained in 'Our approach to developing Codes measures', we consider that applying this and other measures to large, low-risk services will have benefits for users. In particular, we consider that there is a potential material benefit in applying this measure to providers of large, low-risk services because these services are typically complex and are likely to need to moderate large volumes of content (even if the risk of illegal harms is low). Any content moderation failures on such services have the potential to affect a large number of users, with the potential amount of harm being greater as a result. Therefore, it is important that the content moderation functions for these services are adequately resourced. This measure is linked to the measures relating to internal content policies, performance targets, and having a policy for the prioritisation of content for review, and plays a vital role in ensuring these measures have their intended effect. We also consider that providers of large services are likely to have sufficient resources to develop or adjust these policies in line with the proposed measure. We therefore maintain that it is proportionate to apply this measure to providers of large, low-risk services.

2.389 As set out in paragraph 2.356, one stakeholder suggested that this measure should also apply to all single-risk services.[268] Another argued that it should apply to all services (including small, low-risk ones).[269] As explained in paragraph 2.360, this measure is linked to the measures relating to internal content policies, performance targets, and having a policy

[268] VerifyMy response to November 2023 Consultation, p.6.

[269] Barnardo's response to November 2023 Consultation, p.15. We note that the C3P (pp.21-22); Children's Commissioner for England (pp.59-60); Molly Rose Foundation (p.41) and UKSIC (p.36) made a similar point in response to an equivalent measure proposed in the May 2024 Consultation.

for the prioritisation of content for review, and therefore it would be inappropriate to apply it to service providers falling outside the scope of these measures.

2.390 As explained in 'Our approach to developing Codes measures' we are considering whether it may be appropriate to expand the services within the scope of a number of our content moderation measures, including this one, to include single-risk services. We expect to consult again on this in Spring 2025.

2.391 We note that all service providers should ensure that they have adequate resources to enable them to implement the measures relating to reviewing and swiftly taking down content, even where they are given more flexibility as to how they achieve that.

2.392 We are therefore recommending this measure for providers of large services and providers of multi-risk services.

## Conclusion

2.393 Our analysis shows that the measure we are recommending is likely to provide significant protections to users from harm, there will be no additional negative impacts on rights. While the costs could be significant, our analysis suggests that they are proportionate given the scale of the benefits and the fact that resourcing content moderation functions appropriately is critical to protecting people from online harms. Therefore, we have decided to leave the measure unchanged from the measure we proposed in our November 2023 Consultation.

2.394 All providers of large or multi-risk services should resource their content moderation functions to give effect to their internal content policies and performance targets. In doing so, they should have regard to the needs of their UK user base in respect of language as identified in their risk assessments, and the propensity for external events to lead to a significant increase in demand for content moderation on the service. We consider that adequate resourcing of content moderation functions will result in providers making more accurate and timely decisions about whether to remove content.

2.395 This measure will be included in our Codes of Practice for U2U services on Terrorism, CSEA and other duties. It is referred to within these Codes as ICU C6.

# Measure on provision of training and materials to individuals working in moderation (non-volunteers)

2.396 In the November 2023 Consultation, we proposed that people working in content moderation should receive training and materials to enable them to identify and take down illegal content in accordance with their internal content policies. In doing so, we recommended that providers should:

- have regard to at least the risk assessment of their services and information pertaining to the tracking of signals of emerging illegal harm; and
- where providers identify a gap in moderators' understanding of a specific kind of illegal harm, they give training and materials to remedy this.

2.397 Our proposal expressly excluded volunteers. We proposed that this measure should apply to all providers of large U2U services and all providers of multi-risk U2U services.

2.398 In our proposed amendments to the Illegal Content Codes consulted on alongside the May 2024 Consultation, we amended the reference to "emerging harm" to clarify Measure ICU C7 that providers should have regard to "evidence of new and increasing illegal harm on the service (as tracked in accordance with the Measure ICU A5 in Volume 1: chapter 5: 'Governance and accountability' on tracking evidence of new and increasing harm on a service)." [270]

2.399 We explained our view that moderators' ability to effectively carry out their roles in reviewing and taking down content in line with internal content policies is best achieved through training and provision of materials.

## Summary of stakeholder feedback[271]

2.400 In addition to those stakeholders who expressed broader support for the full package of content moderation measures outlined in paragraph 2.14, several stakeholders expressed support specifically for this measure.[272] While not directly referencing this measure, Four Paws said that training is essential for enabling moderators to identify and take down harmful content to protect users.[273] Snap specifically expressed support for the part of the measure recommending that providers remedy gaps in moderators' understanding of illegal harms and this component of the measure's creation of a feedback loop between subject experts, policy developers, and people working in content moderation.[274] Samaritans expressed its support for our recommendation that providers should remedy gaps in moderators' understanding of specific harms. [275]

2.401 Several providers shared information about how they train people working in moderation on their services. Booking.com said its content moderators spend approximately six hours a month receiving training, reviewing content guidelines and policy clarifications, reviewing their errors, and asking questions. It also shared that when new policies are launched, training decks and videos are provided to explain the new content policies and the appropriate actions the moderators should take.[276] Meta also shared strategies it has found helpful in training its moderators. These include providing initial training to review teams to ensure they have a strong grasp of policies, and providing human reviewers who review content alleged to be illegal with specialist training on the particular nature of their work.[277]

[270] We made this amendment to clarify that this measure referred to our existing recommendation that providers should track evidence of new and increasing illegal harm on their services in accordance with the measure in Volume 1: chapter 5: 'Governance and accountability'.

[271] Note this list in not exhaustive, and further responses can be found in Annex 1.

[272] Association of British Insurers response to November 2023 Illegal Harms Consultation, p.3; Born Free Foundation response to November 2023 Consultation, p.5; Cats Protection response to November 2023 Consultation, p.10; Global Network Initiative response to November 2023 Consultation, p.8; Mencap response to November 2023 Consultation, p.8; Meta response to November 2023 Consultation, p.25; Open Rights Group response to November 2023 Consultation, p.2; Snap response to November 2023 Consultation, p.12; Spotify response to November 2023 Consultation, p.4. We note that Meta made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.23.

[273] Four Paws response to November 2023 Illegal Harms Consultation, p.13.

[274] Snap response to November 2023 Consultation, p.12.

[275] Samaritans response to the November 2023 Illegal Harms Consultation, p.3. We note Samaritans made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.8.

[276] Booking.com response to November 2023 Illegal Harms Consultation, p.9.

[277] Meta response to November 2023 Consultation, p.26.

2.402 There were several areas where stakeholders felt this measure should be amended. These included (but were not limited to):

- who receives training;
- frequency of training;
- what training should include;
- who delivers training;
- the costs of this measure; and
- who this measure applies to.

### Who receives training

2.403 Some stakeholders suggested that volunteers should not only be provided with materials (as outlined in the measure below), but also receive training.[278]

2.404 In response to a corresponding measure proposed in the May 2024 Consultation on Protecting Children from Harms Online, the NSPCC recommended that we strengthen the measures we recommend for providers on volunteers, beyond the provision of training materials.[279]

2.405 We address this in paragraph 2.427 under the section entitled 'How this measure works'.

### Frequency of training

2.406 Global Partners Digital and C3P suggested that people working in moderation should receive ongoing training on any changes made to content policies.[280] Furthermore, C3P said that there must be ongoing training provided to update trainees on evolving tactics offenders use to harm children.[281] Meta said that training and testing review teams as appropriate beyond initial training is a strategy it has found helpful.[282]

2.407 We address this in paragraph 2.431 under the section entitled 'How this measure works'.

### What training should include

2.408 Two stakeholders said that people working in moderation should receive training on a service's terms of service or content policies, and how to implement them.[283] Meta specifically argued that training should be proportionate to achieve its purpose, and that if people working in moderation aim to review content for violation of content policies, then they should receive training on such policies.[284]

[278] C3P response to May 2024 Consultation, p.22; Marie Collins Foundation response to November 2023 Consultation, p.10; Meta response to May 2024 Consultation on Protecting Children from Harms Online, p.24.
[279] NSPCC response to the May 2024 Consultation on Protecting Children from Harms Online, pp.51-52.
[280] C3P response to November 2023 Consultation, pp.15-16; Global Partners Digital response to November 2023 Illegal Harms Consultation, p.13.
[281] C3P response to November 2023 Consultation, p.16.
[282] Meta response to November 2023 Consultation, p.26.
[283] Global Partners Digital response to November 2023 Consultation, p.13; Meta response to the 2023 Consultation, p.25. We note that Meta made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.24.
[284] Meta response to the November 2023 Consultation, p.25. We note that Meta made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.23.

2.409 Refuge made a general argument that we should give further direction about what training and materials in the measure should include, including recommending what types of harm should be covered.[285]

2.410 Several civil society organisations recommended specific harms training, including:

- antisemitism (Board of Deputies of British Jews); [286]
- self-harm and suicide content (Samaritans); [287]
- open-source intelligence techniques (Logically); [288]
- risks to child safety (5Rights Foundation); [289]
- assessing and inferring age (Canadian Center for Child Protection); [290]
- the welfare and consumer risks relating to pet sales (Cats Protection); [291]
- training in escalating cases to more senior decision-makers (Global Partners Digital); [292]
- examples of terrorism and hate in the context of Northern Ireland (South East Fermanagh Foundation); and [293]
- training on slang, "harmful meanings for innocuous words" and terms used in different languages which may be harmful (Refuge).[294]

2.411 In response to the May 2024 Consultation on Protecting Children from Harms Online, the Scottish Government argued that an equivalent measure on training should include training on recognising violence against women and girls.[295]

2.412 Some stakeholders commented on the specific skills and qualifications people working in moderation should have.[296] Several suggested that people working in moderation should be able to interpret the cultural, political and social context, as well as the language of content.[297]

2.413 We address these points in paragraphs 2.439 – 2.443 in the section entitled 'Benefits and effectiveness'.

[285] Refuge response to November 2023 Consultation, p.12.
[286] Board of Deputies of British Jews response to November 2023 Illegal Harms Consultation, p.3.
[287] Samaritans response to November 2023 Consultation, p.3.
[288] Logically response to November 2023 Consultation, p.19.
[289] 5Rights Foundation response to November 2023 Consultation, p.21.
[290] C3P response to November 2023 Consultation, p.32.
[291] Cats Protection response to November 2023 Consultation, p.10.
[292] Global Partners Digital response to November 2023 Consultation, p.13.
[293] South East Fermanagh Foundation response to November 2023 Consultation, pp.9-10.
[294] Refuge response to November 2023 Illegal Harms Consultation, p.12.
[295] Scottish Government response to the May 2024 Consultation on Protecting Children from Harms Online, p.15.
[296] SPRITE+ (School of Journalism, Media and Communication, University of Sheffield) response to November 2023 Consultation, p.24; Open Rights Group response to November 2023 Consultation, p.2.
[297] BILETA response to November 2023 Consultation, p.10; Electronic Frontier Foundation response to November 2023 Consultation, p.10; Open Rights Group response to November 2023 Consultation, p.2.

### Who delivers training

2.414 Some civil society stakeholders suggested that providers should collaborate with those with expertise on specific harms to deliver training.[298]

2.415 We address these points in paragraph 2.442 under the section entitled 'Benefits and effectiveness'.

### Costs of the measure

2.416 Snap argued that our estimated costs for the measure omitted the costs of providers giving real-time guidance and support to moderators.[299]

2.417 We address these points in paragraph 2.448 under the section entitled 'Costs and risks'.

### Who this measure applies to

2.418 Several stakeholders argued that this measure should apply to all service providers.[300]

2.419 Some stakeholders argued the measure should apply to single-risk service providers (as well as multi-risk and large service providers).[301]

2.420 We address these points in paragraphs 2.457 to 2.459 under the section entitled 'Who this measure applies to'.

## Our decision

2.421 We have decided to recommend the measure broadly as we proposed in the November 2023 Consultation, except for some small amendments:

- Our measure now says that providers must provide training to "individuals working in moderation" to enable them to "fulfil their roles in moderating content" instead of "to moderate content". It also now says that this recommendation is "including in relation to" instead of "in accordance with" the measures on reviewing, assessing and swiftly taking down content and their internal content policies. These amendments are to acknowledge that individuals working in moderation could have roles in the wider ecosystem of content moderation and may not be directly moderating content themselves.
- In our May 2024 Consultation, we also consulted on an additional measure on the provision of materials to volunteers. We outline this measure in detail in the section below entitled 'Measure on provision of materials to volunteers'.

[298] Refuge response to November 2023 Consultation, p.12; Samaritans response to the November Consultation, p.5. We note that Blue Cross made a similar pint in response to the August 2024 Consultation on Animal Cruelty, p.6. We note that Samaritans made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.8.

[299] Snap response to November 2023 Consultation, p.12.

[300] Barnardo's response to November 2023 Consultation, p.15; Board of Deputies of British Jews response to the November 2023 Consultation, p.3; BT response to November 2023 Consultation, p.2; C3P response to November 2023 Consultation, p.15; Logically response to November 2023 Consultation, p.19. We note that the C3P (p.21-22), Children's Commissioner for England (pp.59-60), Molly Rose Foundation (p.41) and UKSIC (p.36) made a similar point in response to an equivalent measure proposed in the May 2024 Consultation.

[301] Age Verification Providers Association response to November 2023 Consultation, p.2; NSPCC response to the November 2023 Consultation, pp.20-21; VerifyMy response to November 2023 Consultation, p.6; Yoti response to November 2023 Consultation, p.16.

2.422 The full text of this measure can be found in our U2U Illegal Content Codes of Practice and is referred to as ICU C7. This measure will be included in our Codes of Practice on Terrorism, CSEA and other duties.

# Our reasoning

## How this measure works

2.423 We recommend that providers of large services and providers of multi-risk services should ensure individuals working in content moderation (who are not volunteers) receive training and materials that enable them to fulfil their roles in moderating content including in relation to the recommendations in (1) our measures on reviewing, assessing and taking down content swiftly and (2) their internal content policies.

2.424 In the 2022 Illegal Harms Call for Evidence, several service providers told us they train individuals working in moderation to remove illegal (or violative) content and outlined (at a high-level) what kinds of training and support they receive.[302] For example, some service providers told us that new hires in content moderation teams receive onboarding training before commencing their specific roles, which can include training on specific policies, shadowing senior staff to understand how policies and procedures are applied in practice, and training on relevant systems.[303] These service providers also noted that they have on-going training, learning, and development in place, and that performance is assessed via exams. This aligns with the evidence providers shared with us in the November 2023 Consultation about the existing techniques they use to train people working in moderation.[304]

### Individuals working in content moderation

2.425 We expect the individuals working in content moderation would mostly be content moderators employed or contracted by providers, though it could include those who are involved in the wider content moderation ecosystem. This includes, but is not limited to:

- Individuals working on processing appeals;
- trust and safety staff;
- quality assurance and compliance staff;
- subject matter experts;
- lawyers and other legal staff;

[302] Airbnb response to 2022 Ofcom Call for Evidence: First phase of online safety regulation; Nextdoor response to 2022 Ofcom Call for Evidence: First phase of online safety regulation; OnlyFans response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.; Roblox response to 2022 Ofcom Call for Evidence: First phase of online safety regulation; Snap response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.; [✂]; Twitter response to 2022 Ofcom Call for Evidence: First phase of online safety regulation. We note that some providers have also published information on this. Pornhub, 2021. Pornhub Sets Standard for Safety and Security Policies Across Tech and Social Media; Announces Industry-Leading Measures for Verification, Moderation and Detection. [accessed 26 November 2023]; Meta, 2022. How review teams are trained. [accessed25 November 2024].

[303] [✂].

[304] Booking.com response to the November 2023 Consultation, p. 9; Meta response to the November 2023 Consultation, p.26.

- risk management staff;
- operations staff;
- engineers;
- developers.

2.426 We have made a slight amendment to the measure to reflect that, due to the diversity of staff who could be working in moderation, training should be provided to allow them to fulfil their roles including in relation to the measures on reviewing, assessing and taking down content and in internal content policies. These roles might not be restricted to directly moderating content.

2.427 We have considered the inclusion of volunteers in this measure and note that volunteers receive training on some services.[305] However, some service providers use large numbers of volunteers to help them moderate content. Given this and the costs of training per content moderator, we are not currently in a position to determine when, and for which providers, this significant cost burden is justified by the benefit. Therefore, the measure does not recommend that providers train volunteers, although we welcome it where providers do so. We have a separate measure to ensure moderators still have access to appropriate materials, which we outline from paragraph 2.463.

### Materials

2.428 We recommend that in addition to training, providers should also give materials to individuals working in moderation to enable them to fulfil their roles.

2.429 Specific materials provided to content moderators may include the standards about content that falls under the measure on internal content policies as well as any other associated materials. They may also include definitions and explanations around specific parts of the policy, enforcement guidelines, examples, and visuals of the review interface (the tool or interface moderation staff will use to carry out their role).[306]

2.430 Due to the diversity of providers in scope of the measure, we are not being prescriptive about what materials we would expect to be provided to people working in moderation. What is provided may vary depending on a number of factors, such as the type of service, the type of content being moderated, and the local laws and regulations of the region where the service operates.

### Frequency of training

2.431 Some stakeholders emphasised the need for regular updates to training materials and redelivery of training to ensure moderators apply evolving policies consistently. While we acknowledge the benefits of regularly updating training and materials, we are not specifying how frequently this should take place.[307] This is because we have found no set

[305] Discord, no date. Discord Moderator Academy. [accessed 25 November 2024]; Freecycle, no date. Moderator Resources. [accessed 25 November 2024]; Nextdoor, no date. About Review Team members and moderation. [accessed 25 November 2024]; Reddit, no date. Moderator Help. [accessed 25 November 2024]. Twitch, no date. Guide for Moderators. [accessed 25 November 2024]; WhatsApp, no date. 101: Building a Safe Community. [accessed 25 November 2024].

[306] Trust & Safety Professional Association, no date. Setting Up A Content Moderator for Success. [accessed 25 November 2024].

[307] C3P response to November 2023 Consultation, p.16; Global Partners Digital response to November 2023 Consultation, p.13; Meta response to November 2023 Illegal Harms Consultation, p.26.

best practice on how often training or supporting materials should be refreshed.[308] Appropriate frequency may depend on a number of factors, including a person's role and performance, the risks of illegal harm a service faces, and the extent to which such risks vary over time. Therefore, we do not consider that it would be appropriate to specify in Codes how often materials should be revised or training should be repeated.

2.432 However, a provider which failed to refresh training and materials following any major changes to policies or processes relating to content moderation of suspected illegal content or proxy content would not be enabling its moderators to moderate content in accordance with the measures on reviewing, assessing and taking down content swiftly. Furthermore, a provider would not be acting in accordance with this measure if it failed to give training and materials to remedy gaps in moderation staff's understanding of a specific kind of illegal harm, which may include the evolution of tactics offenders use to harm children, as mentioned by C3P.[309] We explain this in more detail in paragraph 3.440.

#### Having regard to risk assessments and new and increasing harm on a service

2.433 In training individuals working in content moderation, we recommend that providers should ensure they have had regard to at least the risk assessment of the service and evidence of new and increasing illegal harm on the service (in accordance with the Measure ICU A5 on tracking evidence of new and increasing harm on a service as outlined in Volume 1: chapter 5: 'Governance and accountability').

#### Remedying gaps in the understanding of individuals working in content moderation in relation to specific kinds of illegal harm through training

2.434 Providers should ensure that where they identify a gap in the understanding of individuals working in content moderation in relation to a specific kind of illegal harm, they give training and materials to remedy this.

### Benefits and effectiveness

2.435 We consider that this measure will deliver significant benefits. This view is reinforced by the support we received for the measure in the November 2023 Consultation, as well as stakeholders' feedback on the importance of training for individuals working in moderation in responses to the 2022 Illegal Harms Call for Evidence and the 2023 Call for Evidence on

[308] We note that Roblox and X (formally known as Twitter) have told us they trained their staff regularly but did not tell us exactly how often they train staff involved in moderation.  Roblox response to 2022 Ofcom Call for Evidence: First phase of online safety regulation; Twitter response to 2023 Call for Evidence: Second phase of online safety regulation. The Trust & Safety Professional Association states on its website that before launching a policy change, staff involved in content moderation need to be trained on the change. Trust & Safety Professional Association, no date. Setting Up A Content Moderator for Success. [accessed 25 November 2024].

[309] C3P response to November 2023 Consultation, p.15.

Protecting Children Online.[310] [311] [312] The importance of training is also supported by broader academic and civil society literature and research.[313]

2.436 Individuals working in content moderation that have been trained on how to identify illegal or violative content are more likely to be equipped with the knowledge and skills to do so effectively (in comparison with untrained individuals). They will be better able to make accurate decisions as to whether content is illegal or violative of a provider's terms of service (depending on how providers choose to review suspected illegal content as outlined in the measure on reviewing and assessing content). This should contribute to reducing the amount of illegal content that remains on a service after review (or does not have any other appropriate action taken on it), and the amount of non-illegal content which is falsely taken down or falsely has other action taken on it. We consider that the measure will both reduce users' exposure to illegal content and help safeguard freedom of expression (by reducing over-removal), and therefore delivers important benefits.

### Having regard to risk assessments and evidence of new and increasing harm on a service

2.437 A service provider's illegal content risk assessment will be one of the key sources of information telling a provider what risk of illegal content it has on its service and will form the basis for internal content policies. As individuals working in moderation should be focused on enforcing the internal content policies, it makes logical sense for training to be informed by the most recent illegal content risk assessment. This will enable providers to ensure training focuses in particular on illegal harms that are most likely to occur on their services and that pose the biggest threat to users. This will make content moderation functions better able to respond to these harms to protect users.

2.438 In Volume 1: chapter 5: 'Governance and accountability', we recommend that providers should track signals of new and increasing illegal harm.[314] This is one of the key sources of

[310] Association of British Insurers response to November 2023 Consultation, p.3; Cats Protection response to November 2023 Consultation, p.10; Spotify response to November 2023 Consultation, p.4; Mencap response to November 2023 Consultation, p.8; Born Free Foundation response to November 2023 Consultation, p.5; Global Network Initiative response to November 2023 Consultation, p.8; Snap response to November 2023 Consultation, p.12; Meta response to November 2023 Consultation, p.25; Open Rights Group response to the 2023 Consultation, p.2.. We note that Meta (p.23) and Samaritans (p.8) supported an equivalent measure in response to the May 2024 Consultation on Protecting Children from Harms Online, p.8.

[311] 5Rights Foundation response to 2022 Ofcom Call for Evidence: First phase of online safety regulation ; Carnegie UK response to 2022 Ofcom Call for Evidence: First phase of online safety regulation ; Center for Countering Digital Hate (CCDH) response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation; Refuge response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation; Glitch response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation ; Global Partners Digital response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation; NSPCC response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation.

[312] 5Rights Foundation response to 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation; Refuge response to 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation; Glitch response to 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation; Global Partners Digital (GPD) response to 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation; SWGfL response to 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation; Samaritans response to 2023 Ofcom Call for Evidence: Second Phase of Online Safety Regulation.

[313] Ofcom, 2019. USE OF AI IN ONLINE CONTENT MODERATION [accessed 25 November 2024]; The Alan Turing Institute, 2021. Understanding online hate: VSP Regulation and the broader context. [accessed 25 November 2024].

[314] Measure ICU A5 in the Illegal Content Codes of Practice.

information on how illegal content manifests and it is therefore crucial providers use this to keep their content moderation training and supporting materials up to date.

### Remedying gaps in the understanding of individuals working in content moderation in relation to specific kinds of illegal harm through training

2.439 In response to both the 2022 Illegal Harms Call for Evidence, and the November 2023 Consultation, several stakeholders commented that this measure should include reference to training on specific types of harms.[315] [316]

2.440 There may be occasions where harms-specific training and materials can be helpful in identifying and removing illegal content due to the unique, complex, novel, or serious nature of a given harm, or because certain harm or harms may be particularly prevalent on a service and so require more in-depth understanding.[317] For example, although some CSAM can be easily identified as illegal content, this is not always the case. Without such training and materials, individuals working in moderation may lack the understanding of a specific illegal harm to be able to correctly review, assess and take down content amounting to such harm. They may also incorrectly take down content that does not amount to illegal harm. Therefore, we consider that remedying gaps in the understanding of individuals working in content moderation in relation to specific kinds of illegal harms through training and materials provides important mitigations against the under-removal and over-removal of content. This is beneficial for protecting users from harm, as well as protecting their freedom of expression rights.

2.441 We note stakeholder points about the importance of individuals working in moderation having an understanding of the cultural, social, and political context of content in order to moderate it effectively.[318] Where the understanding of specific harms depends on the

[315] In the 2022 Ofcom Call for Evidence: First phase of online safety regulation.  , these included tech abuse and gender-based violence (Glitch and Refuge responses to the 2022 Ofcom Call for Evidence: First phase of online safety regulation); child safeguarding, risks to children, and knowledge of child development (5Rights Foundation and NSPCC responses to the 2022 Ofcom Call for Evidence: First phase of online safety regulation); and awareness of learning disabilities (MENCAP response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation). Global Partners Digital also stressed the importance of training moderators in the potential impact to users' rights and freedom of expression (Global Partners Digital response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation).

[316] 5Rights Foundation response to November 2023 Consultation, p.21; C3P response to November 2023 Consultation, p.32; Cats Protection response to November 2023 Consultation, p.10; Board of Deputies of British Jews response to November 2023 Consultation, p.3; Global Partners Digital response to November 2023 Consultation, p.13; Samaritans response to November 2023 Consultation, p.3; South East Fermanagh Foundation response to November 2023 Consultation, pp.9-10; Logically response to November 2023 Consultation, p.19. We note the Scottish Government made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.15.

[317] We note that in response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation, a few services discussed specialist training, including for specific harms. For example, OnlyFans said it had rolled out company-wide mandatory modern slavery and human trafficking training to prevent, detect and report these harms on its service. OnlyFans, 2022. OnlyFans response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation. Nextdoor said that while volunteer community moderators reviewed most types of 'guideline-violating content' on its platform, trained staff handled misinformation and discrimination moderation activities. Nextdoor, 2022. Nextdoor response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation. We also know that many providers, particularly providers of larger services, give their staff involved in moderation specialist training and materials in particular areas, including illegal harms, other harms, freedom of expression, and user rights. Ofcom VSP information gathering from TikTok – 25/07/2022.

[318] BILETA response to November 2023 Consultation, p.10; Electronic Frontier Foundation response to November 2023 Consultation, p.10; Open Rights Group response to November 2023 Consultation, p.2.

individuals working in moderation's understanding of the relevant context of the harm, the provider should remedy any gaps identified in this contextual understanding.

2.442 However, we are not taking forward stakeholder recommendations to be more prescriptive about what harms training and materials should include, recommend providers use external expertise to deliver training, or specify what specific qualifications and skills people working in moderation should have.[319] The measure covers a broad range of services, and different harms will manifest to varied extents and in different ways on them. Whether it is beneficial for providers to use external expertise on harms to deliver training will depend on the harms on a service and how they manifest, and the internal expertise within the service. As stated above, at this stage we consider providers best placed to determine what is appropriate for their services, and the specific harms that may manifest on them.

2.443 For the same reasons, we consider that providers are best placed to determine what skills or qualifications their moderation teams should have. We address the points made around language skills in paragraphs 2.374 to 2.379 in the measure on resourcing.

## Costs and risks

2.444 We consider the main factors driving the cost of training as recommended in this measure to be the number of individuals to be trained and the duration of the training.

2.445 The duration of the training is likely to be longer where there is a more complex and diverse range of possible illegal content on a service. In our November 2023 Consultation, we estimated a duration of two to six weeks for someone receiving training for the first time.[320]

2.446 Based on this duration and a range for pay, we estimate the costs of providing training for one new content moderator to be between £3,000 and £18,000, while the costs for training a new software engineer are estimated to be between £5,000 and £28,000.[321] If content moderators are based in countries with lower labour costs than the UK, the lower end of the assumed wage range may overstate the costs. Costs may also vary depending on whether the training is given by in-house staff or by an external provider.

[319] Open Rights Group response to November 2023 Consultation, p.2; SPRITE+ (School of Journalism, Media and Communication, University of Sheffield) response to November 2023 Consultation, p.24; Refuge response to November 2023 Consultation, p.12; Samaritans response to November 2023 Consultation, p.5. We note Blue Cross made a similar point in response to the August 2024 Consultation on Animal Cruelty, p.7. We note that Samaritans made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.8.

[320] This range is consistent with examples we are aware of from the industry, although in most cases the training requirement is likely to be shorter than six weeks. These estimates are for one-off training, although providers may also provide some refresher training to their employees from time to time, which is likely to vary by service and depend on several factors, including the individual and their role.

[321] This is based on our assumptions on wage rates as set out in Annex 5. We also assume that the wage cost of the people being trained represents only half of the total costs of the training. Other costs included preparing the training materials, running the training and any related travel to the training. This is consistent with the Department for Education saying that the wage cost of the people being trained accounted for about half of all training expenditure in 2019, although this varies by size of firm and sector. Department for Education, 2019. Employer skills survey 2019: Training and Workforce Development - research report ,pp 38 and 40. [accessed 25 November 2024]. Note that the cost estimate for this measure in the November 2023 Consultation excluded the 22% uplift that we have assumed elsewhere for non-wage labour costs, but we have included this in this updated estimate due to a better understanding of the data. We have also updated these figures since the November 2023 Consultation in line with the latest wage data released by ONS.

2.447 In addition to the costs of training new content moderators and software engineers, there will also be some ongoing costs for refresher training and training in new harms emerging on services. We expect that the annual costs of these would be lower.

2.448 Providers whose content moderators require a high level of real-time guidance and support may face higher costs than those outlined in paragraph 2.446. This will tend to indicate that these services are higher risk with more complex content moderation requirements. The benefits from individuals working in content moderation's better understanding the nature of the illegal harms are also likely to be higher in such cases, meaning the costs and benefits will tend to rise together.

2.449 In general, providers of smaller services will have less content to review, and smaller content moderation teams, meaning it is likely that they will incur lower costs. While costs relating to smaller (and larger) services will scale with the risk of harm, this would come with a commensurate benefit. We would generally expect costs to vary with the potential benefits (in the sense that the greater the number of individuals working in content moderation required, the more illegal content tends to be on a service).

2.450 These costs are also mitigated by the fact that this measure does not specify exactly how providers should train individuals working in content moderation, giving them some flexibility in what they do. Providers can decide the most appropriate and proportionate approach to training content moderators for their own contexts. This flexibility provides a cost-effective and proportionate approach for each provider.

2.451 Any additional materials (beyond those used in the initial training) may also incur costs. We do not anticipate that the costs of preparing and producing such materials would add much to the costs of the training.

### Rights impact

2.452 This measure should be seen as part of a package of measures relating to content moderation for illegal content, including the measures on reviewing, assessing and swiftly taking down content, the measure on internal content policies, the measure on performance targets and the measure on a policy for the prioritisation of content for review (for which we have assessed the rights impacts in the relevant sections of this chapter). We do not consider that the measure will have any additional negative impact on users' rights.

2.453 Appropriately training individuals working in content moderation is likely to have significant positive impacts on users' rights, because mistakes are less likely, and moderators will understand their privacy and data protection obligations, where relevant. To the extent that it helps to reduce harm on the service and make users feel safer, this could also positively impact on their human rights. Overall, and taking the benefits to users and affected persons into consideration, we consider that any impact on rights from this measure is proportionate.

### Who this measure applies to

2.454 This measure is linked to, and would be effective for, those service providers which have content moderation policies in compliance with the relevant measure (internal content policies). It follows that it should only be considered for service providers within scope of that measure (providers of large services or multi-risk services). We consider it very unlikely that it would be possible for individuals working in moderation on these services to review content in line with content moderation policies and take action (where appropriate) without training and additional materials (such as definitions and explanations around specific parts of the content moderation policy, enforcement guidelines, examples, and visuals of the tool or interface individuals working in moderation will use to carry out their job).

2.455 Although the additional costs of this measure may be significant for some providers of multi-risk services, we expect the benefits to be high. Training for individuals working in content moderation is important for effectively implementing a service provider's content moderation policies, enabling it to reduce harm and comply with its online safety duties. Well-trained individuals are more likely to be able to identify and action (where required) content in accordance with the measures relating to reviewing and swiftly taking down content, and in accordance with the provider's content standards (as required by the measure relating to internal content policies). As the number of individuals working in content moderation requiring training is likely to depend on the volume of content that needs to be assessed, the costs of this measure are likely to scale with the benefits. We therefore maintain that this measure is proportionate for providers of multi-risk services.

2.456 As explained in 'Our approach to developing Codes measures' we consider that applying this and other measures to large, low-risk services will have important benefits for users as these services have the potential to affect many users and the nature of illegal content can change over time. In particular, we note that large services (even if low risk) are typically more complex, may have a greater volume of content to moderate and more individuals working in content moderation. Appropriate training will therefore reduce the risk of content moderation failures which could affect a large number of users. It will also promote consistency of moderation decisions. Providers of large services are also more likely to have sufficient resources to individuals working in moderation in line with this measure.

2.457 As set out in paragraphs 2.418 and 2.419, a number of responses argued that the measure should apply to all providers of U2U services or to providers of all U2U services with a specific risk.[322] [323] We do not consider it proportionate to apply this measure to providers of small-low risk services as the benefits of requiring them to train individuals working in moderation would be limited due to the low risk of harms. Moreover, as explained in paragraph 2.423 this measure is linked to the measure relating to internal content policies and it would not be appropriate to apply it to service providers who are not within scope of the latter measure.

2.458 As set out in paragraph 2.156, we are considering whether it may be appropriate to apply the measure relating to internal content policies to some smaller, single-risk service providers and will consult on this in Spring 2025. As part of this work, we will also consider whether it may be proportionate to also apply this measure to these service providers.

[322] Barnardo's response to November 2023 Consultation, p.15; BT response to November 2023 Illegal Harms Consultation, pp.1-2; C3P response to November 2023 Consultation, p.15; Logically response to November 2023 Consultation, p.19; Board of Deputies of British Jews response to November 2023 Consultation, p.3. We note that C3P (pp.21-22), Children's Commissioner for England (p.59-60), Molly Rose Foundation (p.41) and UKSIC (p.36) made a similar point in response to an equivalent measure proposed in the May 2024 Consultation on Protecting Children from Harms Online.

[323] Age Verification Providers Association response to November 2023 Consultation, p.2; VerifyMy response to November 2023 Consultation, p.6; NSPCC response to November 2023 Consultation, pp.20-21; Yoti response to November 2023 Consultation, p.16.

2.459 We are therefore recommending this measure for all providers of large U2U services and all providers of multi-risk U2U services.

## Conclusion

2.460 Multi-risk services and large U2U services are not likely to be able to moderate content effectively unless their teams are properly trained. Given the foundational importance of content moderation as a means of protecting users from harm, we therefore consider that the measure in question is proportionate despite the potentially significant costs. Therefore, we have decided to leave the measure largely unchanged from the measure we proposed in our May 2024 Consultation, except for some small amendments to say that providers must provide training to "individuals working in content moderation" to "fulfil their roles in moderating content" instead of "to moderate content". The measure also now says that this recommendation is "including in relation to" instead of "in accordance with" the measures on reviewing, assessing and swiftly taking down content and their internal content policies.

2.461 All providers of large or multi-risk U2U services should provide training and materials to individuals working in content moderation to enable them to fulfil their roles, including in relation to our recommendations on reviewing, assessing and taking down content swiftly and on internal content policies.

2.462 This measure will be included in our Codes of Practice on Terrorism, CSEA and other duties. It is referred to within these Codes as ICU C7.

# Measure on providing materials to volunteers

2.463 In the May 2024 Consultation, we proposed an additional measure that volunteers in relation to content moderation should be provided with materials for their roles. We proposed that this measure should apply to all providers of large U2U services and all providers of multi-risk U2U services.

2.464 We considered that where content moderation volunteers are provided with such materials, they would be more able to carry out their roles in reviewing or reviewing and taking down content swiftly and in accordance with internal content policies.

## Summary of stakeholder feedback[324]

2.465 In addition to those stakeholders who expressed broader support for the full package of content moderation measures outlined in paragraph 2.14, several stakeholders expressed support specifically for this measure.[325]

2.466 Some responses to our November 2023 Consultation referenced the use of volunteer moderation by some services more generally. Reddit referenced its own use of community

[324] Note this list in not exhaustive, and further responses can be found in Annex 1.

[325] Celcis response to the May 2024 Consultation on Protecting Children from Harms Online, p.13; Children's Commissioner for England response to the May 2024 Consultation on Protecting Children from Harms Online, p.60;Kooth response to the May 2024 Consultation on Protecting Children from Harms Online, p.11; Rephrain response to the May 2024 Consultation on Protecting Children from Harms Online, p.16; NSPCC response to the May 2024 Consultation on Protecting Children from Harms Online, p.52; Scottish Government response to the May 2024 Consultation on Protecting Children from Harms Online, p.14.

moderation for user safety, stating that "community moderation happens through a layered, community-driven approach… wherein everyone has the ability to vote and self-organise, follow a common set of rules, establish community-specific norms and ultimately share some responsibility for how the platform works.".[326] Wikimedia Foundation said that communities on its services enforce publicly available policies on content themselves.[327] The Mid Size Platform Group also referenced the use of volunteer moderation by providers and said that the effectiveness of this for mitigating risk was not taken into account in our Codes proposed in November 2023.[328]

2.467 Responses to the May 2024 Consultation identified some areas where stakeholders felt the measure could go further. These included concerns about a large-scale reliance on volunteers and about who the measure applies to. We address this feedback in the following sections.

2.468 Other stakeholder feedback on this measure is addressed in Annex 1.

### Concerns about a large-scale reliance on volunteers

2.469 Meta recommended against a large-scale reliance on volunteers to tackle potential harmful content.[329]

2.470 We address this in paragraph 2.484 in the section entitled 'Benefits and effectiveness'.

### Feedback on who this measure applies to

2.471 We note that in response to an equivalent measure proposed in the May 2024 Consultation, some stakeholders argued that it should apply to all service providers.[330]

2.472 We address this in paragraph 2.495 in the section entitled 'Who this measure applies to'.

## Our decision

2.473 We have decided to recommend the measure broadly as we proposed in the May 2024 Consultation, except for some small amendments.

2.474 Our measure now says that providers must provide materials to volunteers in its content moderation function to enable them to "fulfil their roles in moderating content" instead of "to moderate content". It also now says that this recommendation is "including in relation to" instead of "in accordance with" the measures on reviewing, assessing and swiftly taking down content and their internal content policies. These amendments are to acknowledge that content moderation volunteers could have roles in the wider ecosystem of content moderation and may not be directly moderating content themselves.

2.475 We have decided to proceed with the measure as proposed in the May 2024 Consultation. The full text of the measure can be found in our U2U Illegal Content Codes of Practice and will be included in our Codes of Practice for U2U services on Terrorism, CSEA and other duties. It is referred to as ICU C8.

[326] Reddit response to November 2023 Consultation, p.21-22.

[327] Wikimedia Foundation response to November 2023 Consultation, p.21.

[328] Mid Size Platform Group response to November 2023 Consultation, p.3.

[329] Meta response to May 2024 Consultation on Protecting Children from Harms Online, p.24

[330] C3P response to May 2024 Consultation, pp.21-22; Children's Commissioner for England response to May 2024 Consultation, pp.59-60; UKSIC response to May 2024 Consultation, p.36.

# Our reasoning

## How this measure works

2.476 We recommend that service providers using volunteers in content moderation should ensure that such volunteers are provided with materials that enable them to fulfil their roles in moderating content, including in relation to our recommendations in the measures on reviewing, assessing and taking down content swiftly and the measure on internal content policies. In doing so, the provider should ensure:

a) it has regard to at least the illegal content risk assessment of the service and evidence of new and increasing illegal harm on the service (as tracked in accordance with Measure ICU A5 in Volume 1: chapter 5: 'Governance and accountability'); and

b) where it identifies a gap in such volunteers' understanding of a specific kind of illegal harm, it provides materials to remedy this.

2.477 We are aware that many service providers currently use volunteers – sometimes referred to as community moderators – in content moderation, including illegal content moderation, as reflected in the responses to our November 2023 Consultation about volunteers (see paragraph 2.466). We do not expect that the vast majority of providers to which the measure applies will rely on volunteer moderation alone due to the size of some of the services and the degree of risk of different kinds of illegal content.

2.478 Content moderation volunteers could have a broader role in the wider ecosystem of content moderation, and may not be directly moderating content themselves. To address this, we have amended the measure to recommend that providers must provide materials to volunteers in its content moderation function to enable them to "fulfil their roles in moderating content" and do this "including in relation to the measures on reviewing, assessing and swiftly taking down content and their internal content policies".

2.479 Some providers already offer materials to content moderation volunteers, which provides evidence of the feasibility of this measure.[331] Reddit provides a moderator help centre for its volunteers that includes various courses.[332] Discord offers tools, resources, and guidance as part of its 'Safety Library'.[333] Twitch provides various information pages to aid moderators covering topics such as a 'Guide for Moderators', 'Combating Targeted Attacks', and 'Managing Harassment'.[334] Wikimedia Foundation provides pages on standards requirements.[335]

[332] Reddit provides a moderator help centre containing links to the basics of stating a community on Reddit, overview and explanation of individual moderation tools, community engagement and advice and materials. It also provides sub reddits for news, support and requests. Reddit, no date. Moderator Help. [accessed 25 November 2024]. Reddit also provides volunteers 'Reddit Mod Education Courses'. The layout and set up uses the way that the platform operates to provide materials and support to its users. Reddit, no date. Reddit Mod Education Courses. [accessed 25 November 2024].

[333] Discord provides users information and links to support on its 'safety and moderation' page, including information on how to develop server rules, links to moderation and community support to manage your server and "handling difficult scenarios as an Admin". Discord, no date. Community moderation safety. [accessed 25 November 2024].

[334] Twitch provides various pages that use pictures and videos to show the moderator's view of the channel and its tools. Twitch, no date. Guide for Moderators. [accessed 25 November 2024].

[335] Wikipedia's moderator access is dependent on hierarchy and moderators are required to follow extensive

2.480 As with other measures, we are not being prescriptive about the form these materials should take. We have given providers the flexibility to tailor these resources according to individual needs (provided the content of the resources enables content moderation volunteers to fulfil their role in moderating content, including in relation to the measures on reviewing, assessing, and taking down content swiftly and the measure on internal content policies).

2.481 As we outline in paragraph 2.431, which sets out that paid content moderation teams should be appropriately trained, there is no best set practice on how often materials should be refreshed or updated. However, where there are any major changes to policies or processes relating to content moderation relevant to volunteers, volunteers should be provided with new or updated materials. As with the measure on training individuals working in moderation, where a provider identifies a gap in content moderation volunteers' understanding of a specific kind of illegal harm, it should provide materials to remedy this.

2.482 As outlined in paragraph 2.427, we do not consider that it would be proportionate to recommend that providers should also provide content moderation volunteers with training in addition to materials at this time, due to the significant costs of this recommendation.

## Benefits and effectiveness

2.483 We consider that content moderation volunteers are more likely to carry out their roles effectively if they have access to appropriate materials. Providing such volunteers with materials that enable them to fulfil their role will help them assist with the review and where appropriate swift takedown of content.[336]

2.484 We do not consider that this measure encourages a large-scale reliance on volunteers to tackle harmful content, as implied by Meta.[337] However, we are aware that many service providers currently use volunteers to moderate content and that volunteers on these services often perform a significant proportion of moderation action.[338] In these cases, we

procedural rules. Wikimedia Foundation, 2020. How Content Moderation and Anti-Vandalism Works on Wikipedia.
[accessed 25 November 2024]. Oz, A., 2009. '"Move along now, nothing to see here": The private discussion spheres of Wikipedia', SSRN. [accessed 25 November 2024].

[336] In accordance with Measure ICU C1 and Measure ICU C2.

[337] Meta response to the May 2024 Consultation on Protecting Children from Harms Online, p.24.

[338] Reddit's 2022 Transparency report shows that 58% of content removed from Reddit was actioned by community moderators. The total volume of removals by moderators in 2022 increased by 4.7% compared to 2021. Reddit, 2022.Transparency Report. [accessed 25 November 2024]. In response to the 2022 Ofcom Call for Evidence: First phase of online safety regulation, Mumsnet reported that it has a team of 14 freelance moderators and two staff moderators who are on duty seven days a week. Mumsnet response to 2022 Ofcom Call for Evidence: First phase of online safety regulation. Nextdoor has volunteer moderators on Neighbourhood Teams who monitoring community discussions 24 hours a day, seven days a week. Nextdoor response to 2022 Ofcom Call for Evidence: First phase of online safety regulation. Similarly, Nextdoor's 2022 Transparency Report shows that community moderators reviewed 92% of all reported content. Nextdoor 2022, Transparency Report. [accessed 25 November 2024]. Wikimedia Foundation also uses volunteer moderation, stating that "content moderation on Wikimedia, and other volunteer-run free knowledge projects that the Foundation hosts and supports, is largely conducted by a community of nearly 300,000 global volunteer contributors". It also stated that "the dominant source of moderating and/or governance 'capacity' on the Wikimedia platforms does not come from the service provider at all: it is embodied in the community itself." Wikimedia Foundation response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

consider there will be significant benefits to providing content moderation volunteers materials that enable them to fulfil their roles in content moderation more effectively.

## Costs and risks

2.485 In our May 2024 Consultation, we explained that this measure will incur an initial cost of creating or sourcing materials. This could be done internally (if relevant expertise is available) or externally. Where a provider chooses to source materials externally, the cost will depend on whether it already employs external organisations to provide materials for paid moderators.

2.486 We expect that the majority of providers within the scope of this measure will also be within the scope of the measure relating to the provision of training for individuals working in moderation, and could therefore build on or adapt the materials developed for this measure. There may be small additional costs associated with this (for example, adapting the format of the materials so that they can be accessed online rather than in person, making the materials searchable, or adjusting the level of detail so that the materials are relevant for the role of a volunteer moderator on that particular service). However, we do not anticipate that these costs are likely to exceed a few thousand pounds.

2.487 Costs may be higher for providers that rely solely on content moderation volunteers and do not have relevant existing materials developed for individuals working in moderation. However, we consider that the costs for these providers will be considerably less than the cost estimates for providing training to paid content moderators outlined in the relevant measure on providing training and materials to individuals working in content moderation (see paragraph 2.446). This is because these figures also include wage costs for moderators while receiving the training, which would not be incurred for content moderation volunteers.

2.488 There would also be an ongoing cost to all providers of updating the materials to ensure that they remained relevant. Even if it were possible to source an ‘off the shelf’ version of the materials, this would need to be updated and regularly reviewed in light of new and emerging types of illegal content.

2.489 These costs will be mitigated by the fact that this measure does not specify exactly how materials should be provided to content moderation volunteers, giving providers some flexibility to decide the most appropriate and proportionate approach for their own contexts.

## Rights impact

2.490 This measure should be seen as part of a package of measures relating to content moderation for illegal content, including the measures on reviewing, assessing and swiftly taking down content, the measure on internal content policies, the measure on performance targets, and the measure on a policy for the prioritisation of content for review, for which we have assessed the rights impacts in the relevant sections of this chapter. We do not consider that this measure will have any additional negative impact on users’ rights.

2.491 Providing appropriate materials to content moderation volunteers in a provider’s content moderation function is likely to have significant positive impacts on users’ rights, because mistakes are less likely, and moderators will understand their privacy and data protection obligations, where relevant. To the extent that it helps to reduce harm on the service and make users feel safer, this could also positively impact on their human rights. Overall, and taking the benefits to users and affected persons into consideration, we consider that any impact on rights from this measure is proportionate.

### Who this measure applies to

2.492 We consider that this measure will reduce the risk of illegal content on U2U services by ensuring that content moderation volunteers are better equipped to assist with the identification and swift takedown (where required) of illegal content where it is detected on these services.

2.493 We expect that the costs of implementing this measure for providers of large services or multi-risk services will generally be small because we expect the majority of these services would also be within scope of the measure relating to training and materials for individuals working in moderation. Therefore, they will already have in place materials for individuals working in moderation in most cases, and the additional cost of making some of these materials available to content moderation volunteers should be minimal. The costs of this measure are also likely to be lower for providers of smaller, less complex services with fewer risks because their content moderation volunteers are likely to be dealing with fewer types of illegal content.

2.494 For providers that rely only on content moderation volunteers and do not have existing resources developed for individuals working in moderation that can be adapted, costs will be higher. However, we consider that the benefits from implementing the measure will also be higher, as having well-informed and well-prepared content moderation volunteers will be particularly important where providers place greater reliance on these moderators to reduce the risk of moderation failures.

2.495 Providers of smaller services that are not multi-risk and use content moderation volunteers should consider how to provide relevant information to such volunteers (if appropriate) as part of implementing the measure relating to reviewing and swiftly taking down content. However, we are not necessarily recommending they follow the specific approach to providing materials as described in this measure. The benefits from this measure would be lower on these services given their limited risks of illegal content. The costs of this measure would also be substantial for providers of such services, as they would not be in scope of the measure relating to providing training and materials to individuals working in moderation and therefore may not have existing materials that can be adapted. Therefore, at this time we do not consider it would be proportionate to recommend this measure for providers of such services.

2.496 As set out in 'Our approach to developing Codes measures', we are considering extending a number of measures, including this one, which currently apply to multi-risk U2U services to single-risk U2U services in future. We expect to consult on this again in Spring 2025.

2.497 We are therefore recommending this measure for all providers of large U2U services and all providers of multi-risk U2U services.

## Conclusion

2.498 Our analysis indicates that the measure we are recommending is likely to deliver material benefits. We consider the costs that will result from it are modest and proportionate and that if anything it will have a positive impact on rights. Therefore, we have decided to leave the measure largely unchanged from the measure we proposed in our May 2024

Consultation, except some small amendments so our measure now says that providers must provide materials to volunteers in its content moderation function to enable them to “fulfil their roles in moderating content” instead of “to moderate content”. It also now says that this recommendation is “including in relation to” instead of “in accordance with” the measures on reviewing and swiftly taking down content and their internal content policies.

2.499 All providers of large or multi-risk U2U services should provide materials to volunteers in their content moderation function to enable such volunteers to fulfil their roles in moderating content, including in relation to our recommendations on reviewing, assessing and taking down content swiftly and on internal content policies.

2.500 We consider that where content moderation volunteers are provided materials, they will be able to carry out their roles more effectively in protecting users from illegal harms.

2.501 This measure will be included in our Codes of Practice for U2U services on Terrorism, CSEA and other duties. It is referred to within these Codes as ICU C8.

# 3. Search Moderation

## What is this chapter about?

Search moderation is used by providers to review search content and where relevant, take action to minimise the risk of a wide variety of illegal content from being presented to users in search results, as well as legal content that is covered by their publicly available statement (an illegal content proxy). This chapter sets out the search moderation measures we are recommending, why we are recommending them, and to which search services they should apply.

## What decisions have we made?

We are recommending the following measures:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| ICS C1 | Providers should have systems and processes designed to review, assess and where relevant take **appropriate action** in relation to **illegal content or illegal content proxy** of which they are aware (a 'search moderation function'). | Providers of search services. |
| ICS C2 | Providers should **set and record internal content policies**. | • Providers of large general search services.<br>• Providers of multi-risk search services. |
| ICS C3 | Providers should **set and record performance targets** for their search moderation function. | • Providers of large general search services.<br>• Providers of multi-risk search services. |
| ICS C4 | Providers should prepare and apply a policy in respect of the **prioritisation of search content for review.** | • Providers of large general search services.<br>• Providers of multi-risk search services. |
| ICS C5 | Providers should **resource their search moderation function**, so as to give effect to measure ICS C2 and measure ICS C3. | • Providers of large general search services.<br>• Providers of multi-risk search services. |
| ICS C6 | Providers should ensure **people working in search moderation** (non-volunteers) **receive training and materials** that enable them to **fulfil their role** in moderating search content, including in relation to measure ICS C1 and measure ICS C2 | • Providers of large general search services<br>• Providers of multi-risk search services. |

### Why have we made these decisions?

Effective search moderation systems are able to identify and enable providers to make timely and accurate decisions about search content that is suspected to be illegal, and take appropriate action to protect users. Providers with ineffective search moderation functions may face increased risk of harm on their services.

Our analysis suggests that harm to users will be reduced where providers set content policies, resource and train their search moderation teams appropriately and take into account the likely severity of content and the risk that it will be encountered by a high number of UK users when deciding what content to prioritise for review. We do not think a one-size-fits-all approach to search moderation would be appropriate. Instead of making very specific and prescriptive recommendations about search moderation, we have therefore decided to make a relatively high-level set of recommendations which would allow providers considerable flexibility about how to set up their search moderation teams. We have focussed the most rigorous proposals in this area on providers which are large or multi-risk. This will help ensure that the impact of the measures is proportionate. Similarly, the flexibility built into our proposals will make it easier for providers to carry them out in a way which is cost-effective and proportionate for them.

# Introduction

3.1 In chapter 2 of this Volume: 'Content moderation' we considered proposals in relation to content moderation on user-to-user ('U2U') services. In this chapter, we consider what steps search services should take by way of moderation.

3.2 Search moderation is used by providers to review content and where relevant, take action to minimise the risk of different types of illegal content being presented to users in search results, as well as legal content that is covered by the content policies of the service. While search content policies usually identify illegal content as being subject to moderation action, they do not necessarily closely reflect the requirements of any single legal system due to the global nature of many services.[339]

3.3 Search moderation functions differ between services and are designed to meet the specific needs and contexts of that service. Search moderation can be carried out by humans, automated tools, or a combination of the two. We acknowledge that service providers use a combination of techniques to moderate content and that each moderation system has its own unique benefits and risks.

3.4 The measures within this chapter aim to enable providers to make timely and accurate decisions about search content that is suspected to be illegal and take appropriate action to protect users.

## The Online Safety Act 2023

3.5 Under the Online Safety Act 2023 ('the Act'), search service providers must take steps to minimise the risk of online users encountering illegal search content, and to effectively

[339] Policy Department for Economic, Scientific and Quality of Life Policies, 2020. Online Platforms' Moderation of Illegal Content Online: Laws, Practices and Options for Reform. [accessed 24 November 2024].

mitigate and manage the risks of harm to individuals from illegal content that is accessible in or via their search service.[340]

3.6 As set out in chapter 6 of this Volume: ‘Reporting and complaints’, search service providers also have a duty to respond to complaints about illegal content and to handle appeals from interested persons (i.e. website or database operators) when action is taken in respect of search content because it is identified as illegal.[341] The complaints duties outlined in the Act assume that search services may take or use measures in order to comply with their safety duties in a way that results in content no longer appearing or being given a lower priority in search results.

3.7 Therefore, while the Act does not expressly require search service providers to have a proportionate ‘content moderation’ function, the effect of the safety and complaints duties outlined above is that a moderation function is required in practice in order to minimise the risk of users encountering illegal content in or via search services. This function should be capable of making judgments about whether search content should be treated as illegal content and taking appropriate moderation action where illegal content is identified. We are calling this function ‘search moderation’.

3.8 Search content includes any content that may be encountered in or via search results. Content is to be treated as ‘encountered via’ search results where it is accessed by interacting with search results (for example, by clicking on them) but does not include content encountered via subsequent interactions.[342] In practice, this means that ‘search content’ includes content on a webpage that can be accessed by clicking on a search result (but not content encountered beyond this first interaction). The safety duties outlined in the Act and the measures we recommend for the purposes of complying with them in this chapter should be considered in this context.

## Structure of this chapter

3.9 In the next section, we explain the general approach taken to the search moderation measures. We begin by outlining the approach proposed in our November 2023 Illegal Harms Consultation (‘November 2023 Consultation’), and then explain the approach we have decided to take based on the stakeholder feedback we received.

3.10 Following that, we explain in detail the six measures under this approach.

# Our approach

3.11 As with U2U content moderation, we considered three potential approaches to drafting the search moderation measures in our November 2023 Illegal Harms Consultation (‘November 2023 Consultation’):

- **Approach 1** – specify in detail how providers should configure their search moderation systems and processes.

[340] Section 27(2) and (3) of the Act.
[341] Section 32 of the Act.
[342] Section 57(2) and (5) of the Act.

- **Approach 2** – specify in detail the outcomes search moderation systems and processes should achieve, for example, by setting detailed key performance indicators (KPIs), but leave the design to services.
- **Approach 3** – recommend providers to operate a search moderation system and (where relevant) set out the factors to which they should have regard when designing their search moderation systems and processes.

3.12 We proposed that **Approach 3** would be most appropriate as the evidence suggested that there is no one-size-fits-all approach to search moderation. We recognised that systems and processes for search moderation may differ across service providers and are designed to meet specific needs and contexts. We considered Approach 3 was particularly beneficial given the diverse range of services in scope of the regulation and the fast-moving pace of technological development.

3.13 We proposed not to pursue **Approach 1** or **Approach 2** because:

- we did not have enough evidence to specify in detail every aspect of how providers should configure their search moderation systems and processes, or the outcomes that those systems and processes should achieve;
- there is no consensus on an approach to search moderation;
- different approaches may be more appropriate in different circumstances and for different types of services; and
- taking a prescriptive approach at this stage would give rise to a substantial risk of regulatory failure and unforeseen consequences, which could lead to significant disruption in the sector. We considered that this could lead to potentially increased, rather than decreased, harm to users.

## Feedback on our approach[343]

3.14 There was positive feedback on our approach and the types of measures we proposed based on this approach.[344] Despite one response describing the measures as too broad, other responses from search services supported our position that there is no one-size-fits-all approach to search moderation.[345]

## Decision on our general approach to search moderation

3.15 We have decided to amend the search moderation measures, where necessary, to enable providers to implement them with more flexibility, based on stakeholder feedback outlined in this chapter. This has included small amendments to the measures and specifying the

[343] Note this list in not exhaustive, and further responses can be found in Annex 1.

[344] Are, C. response to November 2023 Illegal Harms Consultation, p.7; Betting and Gaming Council response to November 2023 Illegal Harms Consultation, p.7; Cyber Helpline response to November 2023 Illegal Harms Consultation, p.12; Marie Collins Foundation response to November 2023 Illegal Harms Consultation, p.11; Mencap response to November 2023 Illegal Harms Consultation, p.9; pp.8-9; National Society for the Prevention of Cruelty to Children (NSPCC) response to November 2023 Illegal Harms Consultation, p.23; National Trading Standards eCrime Team response to November 2023 Illegal Harms Consultation, p.8; Nexus response to November 2023 Illegal Harms Consultation, p.10; Stop Scams UK response to November 2023 Illegal Harms Consultation, pp.10-11; Welsh Government response to November 2023 Illegal Harms Consultation, p.3; We Protect Global Alliance response to November 2023 Illegal Harms Consultation, p.13.

[345] Dwyer, D. response to November 2023 Illegal Harms Consultation, p.4. For example, see [✂].

outcomes that appropriate moderation action should achieve, rather than specifying technical actions. We consider that these amendments have addressed concerns that some of the measures were too prescriptive and did not adequately account for the different ways a search service may operate. We outline this feedback and our amendments [in the measure specific sections - reference]. With these amendments, we have adopted what we consider to be a proportionate approach that is appropriate for the different types of services within scope of the duties and our goals relating to improving user safety.

3.16 We consider that the reasoning provided at consultation for not taking a more prescriptive[346] approach for these measures, or an entirely outcomes-based approach for these measures still stand. The reasons for which are set out above in paragraph 3.13.

3.17 We have therefore decided to broadly adopt Approach 3, when designing our search moderation measures. However, consistent with the hybrid approach to designing our measures[347], we have provided more specificity where we consider it appropriate to do so.

## Measure on taking appropriate moderation action in relation to illegal search content

3.18 In the November 2023 Consultation, we proposed that providers should put in place search moderation systems and processes that are designed so that search content that is illegal is deindexed or downranked for UK users.[348] [349] We proposed to recommend that providers have regard to certain factors when considering which moderation action to take (and the extent of that action). These factors included the prevalence of illegal content hosted by the person responsible for the website or database, the interests of users in receiving any lawful material affected by moderation action, and the severity of potential harm of the illegal content. We explain them at paragraph 3.69.

3.19 The proposed measure was designed to reflect the duties in the Act. In effect, these require that search service providers must have in place systems or processes to moderate search content that is illegal content, as explained in paragraphs 3.5 – 3.8.

3.20 We proposed this measure should apply to all search service providers.

### Summary of stakeholder feedback[350]

3.21 Several stakeholders expressed support for the inclusion of this measure in our Codes.[351] Big Brother Watch supported our decision not to recommend deindexing wherever illegal search content is identified, albeit with concerns about the impact of downranking on access to information.[352]

[346] This is particularly the case given the rapid changes taking place in search as generative artificial intelligence (AI) technologies are rolled out.

[347] See ‘Our approach to developing Codes measures’.

[348] See ‘Glossary’.

[349] See ‘Glossary’.

[350] Note this list in not exhaustive, and further responses can be found in Annex 1.

[351] Barnardo’s response to November 2023 Illegal Harms Consultation, p.15; Basra, Dr R. response to November 2023 Illegal Harms Consultation, p.1; Canadian Centre for Child Protection (C3P) response to November 2023 Illegal Harms Consultation, p.17; [✂]; Cyber Helpline response to November 2023 Consultation, p.12; We Protect Global Alliance response to November 2023 Illegal Harms Consultation, p.13.

[352] Big Brother Watch response to November 2023 Illegal Harms Consultation, p.6.

3.22 On the other hand, some civil society organisations called for greater expectations on search services to be more proactive in their approach to moderating search content.[353]

3.23 We also received feedback raising concerns about the technical application of this measure and our assessment of the measure's impact on user rights.[354] We set out these concerns in more detail under the following sub-headings:

- whether actions apply at the URL level or domain level;
- proactive moderation;
- prescribing technical actions;
- interaction with reporting and complaints;
- factors relevant to assessing what appropriate action is; and
- feedback on rights impact.

3.24 We describe the feedback in the following sections, and address it in the 'Our reasoning' section.

### Whether actions apply at the URL level or domain level

3.25 Google asked for our Codes to be explicit that any moderation action would be expected to be taken at the URL level, to avoid the risk of over-removal.[355] We address this request in paragraph 3.52.

### Proactive moderation

3.26 Barnardo's was concerned that that our proposals for search moderation do not include measures related to actively detecting illegal content.[356] The Molly Rose Foundation argued that it was reasonable to expect that platforms should have appropriate ongoing detection and monitoring processes to track emerging changes in user behaviour and search terms, and to apply relevant measures.[357] These points are addressed at 3.50.

### Prescribing technical actions

3.27 Protection Group International and [✂] disagreed that the measure should allow cases where a provider downranks illegal content instead of deindexing it.[358] Additionally, Google argued that providers should not be expected to downrank in circumstances where a URL contains any amount of unlawful content and that providers should have the option to delist.[359] [360]

[353] 5Rights Foundation response to November 2023 Illegal Harms Consultation, p.21; Barnardo's response to November 2023 Consultation, pp.15-16; Christian Action Research and Education (CARE) response to November 2023 Illegal Harms Consultation, p.9; Institute for Strategic Dialogue response to November 2023 Illegal Harms Consultation, p.9; Molly Rose Foundation response to November 2023 Illegal Harms Consultation, pp.35-36.

[354] This feedback comes from responses to our November 2023 Consultation, and, where there is applicability to other Codes measures, responses to our May 2024 Consultation on Protecting Children from Harms Online ('May 2024 Consultation').

[355] Google response to November 2023 Consultation, p.39.

[356] Barnado's response to November 2023 Consultation, p.17.

[357] Molly Rose Foundation response to November 2023 Consultation, p.39.

[358] [✂]; Protection Group International response to November 2023 Illegal Harms Consultation, p.6.

[359] Google response to November 2023 Illegal Harms Consultation, p.39.

[360] See 'Glossary'.

3.28 Skyscanner and Mid Size Platform Group noted that while vertical search services do remove illegal content, the technical actions of ‘deindexing’ or ‘downranking’ do not reflect the fundamentally different way in which these services operate compared to general search services.[361] Both requested further clarity with regard to the wording of this obligation to make clear that other systems or processes that see illegal content removed from search results are also acceptable.

3.29 Google stated that it generally delists illegal content rather than deindexing it, as deindexing does not enable the kind of flexibility that its products rely on and which may be required to comply with other duties in the Act – for example in the event of a successful appeal by a website operator, content that has been blocked through delisting rather than deindexed, can be reinstated into search results immediately as it remains in the underlying index. Further, a deletion from the index has the effect of removing a URL from the overall search index meaning it no longer appears in results anywhere internationally.[362]

3.30 Google also argued there was a lack of clarity in the use of the term “downranking”, given that the same page might rank differently depending on the associated query.[363] In follow up engagement, Google expressed concerns that our proposed measure on appeals in chapter 6 of this Volume: ‘Reporting and complaints’ would bring its entire ranking system within the scope of OSA appeals.[364]

3.31 We explain and address these points in paragraphs 3.57 – 3.59 under the section ‘How this measure works’.

## Factors relevant to assessing what appropriate action is

3.32 Google suggested an amendment that providers should be given the choice as to which moderation action to take, without having to assess the listed factors, and that these should not be prescriptively set out.[365] In response to our May 2024 Consultation, Microsoft noted that a service may not have the context required to judge the severity of potential harm of content on a third-party website.[366] We clarify our position on this in paragraph 3.72.

3.33 Google also argued search services are not able to determine “prevalence” of the illegal content, based on the understanding that “prevalence” is relevant to the presence of illegal content at the broader website or domain. This is because they do not host the site and do not record metrics like violative view rates.[367] We clarify our expectations about undertaking an assessment of prevalence in paragraph 3.73.

3.34 In a response that also affects chapter 2 of this Volume: ‘Content moderation’, the Alliance to Counter Crime described how when prioritising what content to review, “harm to

[361] Mid Size Platform Group response to November 2023 Illegal Harms Consultation, p.8; [✂]; Skyscanner response to November 2023 Illegal Harms Consultation, p.17. Skyscanner made similar points in response to the May 2024 Consultation on Protecting Children from Harms Online, p.14.

[362] Google response to November 2023 Consultation, p.38.

[363] Google response to November 2023 Consultation, p.39. We note that Google raised a similar point in their response to our May 2024 Consultation on Protecting Children from Harms Online, p.28.

[364] Ofcom/Google meeting, 3 October 2024.

[365] Google response to November 2023 Consultation, pp.39-40.

[366] Microsoft response to May 2024 Protecting Children from Harms Online Consultation, p.13.

[367] Google response to November 2023 Consultation, p.39.

children" should be regarded as an aggravating factor when considering severity of content.[368] This is addressed in paragraph 3.73.

### Feedback on rights impact

3.35 The Canadian Centre for Child Protection (C3P) argued that we had not adequately acknowledged the positive impacts of search moderation on the rights of child sexual abuse material (CSAM) victims and survivors.[369]

3.36 The Information Commissioner's Office (ICO) asked for clarity on how personal data is being processed, given that search moderation will likely mostly be dealing with web page content rather than users of a service and their personal data.[370] ICO also queried whether any reporting of CSEA content to the NCA or other law enforcement agencies would have any impact on the rights of search service users and said that it expected the impact to be limited to third party personal data contained on webpages.[371]

3.37 This feedback is addressed in the 'Rights impact' section.

## Our decision

3.38 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation, though we have made changes in response to the feedback set out in paragraphs 3.32 – 3.34.

3.39 We have clarified that the provider's search moderation functions should include systems and processes designed to enable the provider to **review and assess** search content that the provider has reason to suspect may be illegal content, in to taking moderation action.

3.40 We have decided to no longer specify the two technical actions that we recommend service providers take in relation to illegal content. The measure now recommends that where a provider becomes aware of illegal content and/or illegal content proxy,[372] it should take "appropriate moderation action" that results in one of the following outcomes:[373]

- The search content no longer appears in search results for users (for example, because it has been 'delisted' or 'deindexed').
- The search content is given a lower priority in the overall ranking of search results (for example, because it has been 'downranked'). We have clarified, for the avoidance of doubt, that this action does not require illegal search content to appear lower than other search content where this is not possible in response to a given search request, because: 1) only illegal search content is relevant to the user's request, or 2) given the specificity of the request, illegal search content is reasonably considered to be the most relevant response.

[368] Alliance to Counter Crime Online (ACCO) response to November 2023 Illegal Harms Consultation, p.4; a similar point was raised by Hall, J. response to November 2023 Consultation, p.7.

[369] C3P response to November 2023 Consultation, p.17.

[370] Information Commissioner's Office (ICO) response to November 2023 Illegal Harms Consultation, pp.17-18.

[371] ICO response to November 2023 Illegal Harms Consultation, pp.17-18.

[372] In the Codes, we define "illegal content proxy" as content of a kind that is identified in the provider's publicly available statement 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).

[373] The term "appropriate moderation action" is also now used in other provisions of this measure as necessary to replace references to "deindexing" and "downranking".

3.41 We have also made a number of clarificatory amendments in response to stakeholder feedback:

- We have clarified that providers should have regard to the factors in designing the particular aspects of the search moderation function relating to what appropriate moderation action to take (including, where relevant, the extent to which search content is given a lower priority in the overall ranking of search results).
- We have specified how the provider should have regard to the prevalence of illegal content hosted at the URL or in the database where the search content is stored, as the previous drafting could have been perceived as being broader than this.
- We have clarified that by "severity of harmfulness", we mean the severity of potential harm to UK users if they encounter illegal search content on the service. The measure also now explicitly states that potential harm to children is an aspect to be considered as part of the severity of content.

3.42 We have also included additional references to privacy safeguards in this measure to clarify how privacy rights are protected by our measures.

3.43 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C1. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

## Our reasoning

### How this measure works

3.44 We recommend that the provider of a search service should have a search moderation function designed to review, assess and take appropriate moderation action in relation to search content the provider has reason to suspect may be illegal content, to help minimise the risk of users encountering it. This measure applies to all search services.

#### Identifying illegal content

3.45 For this purpose, when a provider has reason to suspect that search content may be illegal content, the provider should either:

a) Make an illegal judgement in relation to the search content; or
b) Assess the search content against the types of content identified in the publicly available statement as being subject to appropriate moderation action. The provider may do this where it is satisfied that the types of content included in the publicly available statement are broad enough to cover the type of illegal content that it suspects exists.

3.46 We also consider it may be appropriate for a search moderation system to adopt an approach that combines the two processes above.

3.47 We have given providers these options because, as outlined in the November 2023 Consultation, we recognise that many service providers design their publicly available statements and community guidelines both to comply with existing laws in multiple jurisdictions and to meet their own commercial needs. We therefore consider that service providers should have a choice. They may set about making illegal content judgements in relation to individual pieces of content for the express purpose of complying with the safety duties. In practice this would necessarily give effect to the publicly available statement the provider adopts under section 27(5) of the Act (which sets out how users are to be protected from illegal content). The alternative is that they moderate illegal content by reference to types of content included in the provisions of their publicly available statement which are cast broadly enough to necessarily cover illegal content.

3.48 Where a provider assesses search content that it suspects to be illegal content against its publicly available statement (rather than making an illegal content judgement), this content would be an "illegal content proxy".

3.49 As set out in our November 2023 Consultation, we consider that providers may be alerted to content they suspect may be illegal content (as the Act defines it) in a variety of ways.

- The Act governs its treatment of complaints by UK users and affected persons, which we consider further in chapter 6: of this volume 'Reporting and complaints'. A complaint by a UK user or affected person about suspected illegal content is grounds to suspect the content may be illegal, except where the provider determines it to be manifestly unfounded as set out from paragraph 6.267 of chapter 6.
- In chapter 6, we also recommend a means for entities with appropriate expertise and information ('trusted flaggers') to report suspected illegal content to service providers. A report from a trusted flagger about matters within its expertise would always be a ground for suspecting the content may be illegal content.
- Additionally, in chapter 5 of this Volume: 'Automated search moderation', we recommend that providers of general search services use automated systems and process to detect URLs at which CSAM is present and remove these from search results. [374]
- Providers may choose to use other kinds of technology or human content moderators in order to identify suspected illegal content as defined in the Act.

3.50 As outlined in paragraph 3.26, some organisations requested to have greater expectations placed on search services to be proactive in their moderation of content.[375] We recognise the value of services undertaking proactive review to detect potentially illegal content before users are exposed to harm. While our measure applies regardless of how providers are alerted to the presence of potentially illegal content, we do not have sufficient evidence at this time about the effectiveness, accuracy or practicality of using existing automated technologies to detect different types of illegal harms on search. Therefore, beyond our measure relating to detection of known CSAM URLs, we do not consider it appropriate at this stage to specifically recommend the use of automated content detection for search moderation. That being said, we welcome steps taken by search providers to deploy automated technologies to meet their duties and encourage ongoing investment and innovation in this area.

[374] We consider that proactive moderation of large volumes of content usually requires the use of automated technologies. We outline our approach to considering these technologies in more detail in chapter 5: 'Automated search moderation'. In that chapter, we recommend the proactive identification and removal of known CSAM URLs from search results in our measure about automated search moderation.

[375] Barnado's response to November 2023 Consultation p.17; Molly Rose Foundation response to November 2023 Consultation, p.39.

### Taking appropriate moderation action

3.51 Where the provider determines that the search content is illegal content, or that it is an illegal content proxy covered by its publicly available statement as outlined above, it should take appropriate moderation action. The appropriate action should mean that the illegal search content either no longer appears in search results or is given a lower priority in the overall ranking of search results.

3.52 We would expect appropriate moderation action to be applied to the illegal content concerned, and so would normally impact at URL level rather than at domain level for general search services, in response to Google's points at 3.25.

3.53 While we have prescribed two outcomes, appropriate moderation action will depend on the way providers operate their search service and the piece of search content in question. We consider that the technical actions proposed in our November 2023 Consultation, deindexing or downranking, may be appropriate technical processes to minimise the risk of users encountering illegal content in certain situations. But we accept the arguments made to us in Consultation responses that they are unsuitable in others. While some stakeholders argued that providers should not have the option to 'downrank' or give illegal search content a lower priority when they should be taking it out of search results, we continue to propose two possible outcomes rather than one.[376]

3.54 A more restrictive position requiring removal of illegal content in all circumstances would not be proportionate where we consider there to be less onerous means of complying with the duties imposed on search services. These duties are to minimise the risk of users encountering illegal content that a provider is aware of, and to mitigate and manage the risk of harm. A decision to remove search content that is found to contain illegal content would render all content at the URL (or equivalent) inaccessible, including legal content hosted on the same page. This would engage the fundamental right to freedom of expression of those that operate the URL or database (as relevant), service providers and users as described in more detail at paragraph 3.93. It could also have detrimental commercial consequences for the operators of URLs and databases. For example, some countries' laws are different from those of the UK. The definition of illegal content means that content from other countries which is lawful in those countries could still be illegal content under the Act. Examples include the priority offences relating to weapons and sex work.

3.55 After careful consideration, we have amended our measure to give providers greater flexibility to use the most appropriate moderation action for how their service operates (provided it achieves one of the two outcomes). It also allows greater flexibility to continue to invest in new technology and in more effective methods for minimising the risk of harm to users from encountering illegal content. The sections below explain in more detail our expectations of the two outcomes.

3.56 In our assessment, moving to an outcomes-based approach for this measure rather than recommending prescriptive technical actions reflects our proportionate, no one-size-fits-all approach to search moderation, that more closely reflects the technical reality of how search services operate.

[376] [✂]; Protection Group International response to November 2023 Consultation, p.6.

### Outcome 1: Illegal search content no longer appearing in search results

3.57 Responses to our November 2023 Consultation explained that the technical action of ‘deindexing’ does not provide adequate flexibility for search providers implementing this measure, and may not be technically feasible in all circumstances, especially for vertical search services.[377]

3.58 We note these concerns and acknowledge there may be circumstances in which ‘deindexing’ as a specific technical function is inappropriate. For example, where content may be illegal within the UK but not in other jurisdictions, deindexing would impact the ability of providers to surface deindexed content to users outside of the UK. We have therefore decided to use more outcomes-based language, recommending instead that providers take action that results in search content no longer appearing in search results (or meet Outcome 2 discussed below). This gives search providers greater flexibility to take whatever action is most technically appropriate for their service or the context. This includes allowing services to delist in situations where deindexing is not appropriate, or, in the case of vertical search services, use other technical means which they may have in place to remove illegal content.

3.59 We are confident that this change will better enable services to pursue improvements in user safety as they are no longer required to use a prescribed technical action, where another action would more effectively achieve the outcome intended.

### Outcome 2: Illegal search content is given a lower priority in overall ranking

3.60 In our November 2023 Consultation, we specified the technical action of ‘downranking’ as one of two actions that we recommend providers take in relation to illegal search content. However, in response to stakeholder feedback, we have decided to recommend that providers take any technical action they consider appropriate that achieves the outcome of illegal search content being given a lower priority in the overall ranking of search results.

3.61 In this sub-section we clarify our expectation of this outcome in the context of a service’s ranking system, before describing what we expect the outcome to look like in practice.

3.62 We understand many general search providers operate their services using complex ranking systems that consider a wide range of factors such as accuracy, authority, or usability, when determining how to prioritise relevant search content in search results. We recognise that these systems are a necessary commercial aspect of search providers’ operations, and that they impact the availability and/or accessibility of certain search content available on or via a search service, including search content that may later be found to be illegal content.

3.63 Due to the complex nature of how providers of general search services index and prioritise search content, we provide further explanation in this section to give greater regulatory certainty about how a provider might achieve this outcome in practice. We also recognise that vertical search services operate in particular contexts, and the range of technical actions available to them may relate to or depend on their specific context. We explain how our change to more outcomes-based language better ensures all search providers can implement this measure in a way that is compatible with their service’s operations.

3.64 We understand that an indication of illegality is just one of a number of parameters considered by providers when ranking content. Google argued that the dependence on the

[377] Skyscanner response to November 2023 Consultation, p.17; Skyscanner made similar points in response to the May 2024 Consultation, p.14.; [✂]; Google response to November 2023 Consultation, p.38.

wording of a user's search request for how search results appear made 'downranking' an unsuitable expectation for the purposes of this measure. It explained that different search results will have different degrees of relevance depending on the specific query entered into the search engine by the user, which in turn will impact the ranking of results. Furthermore, where queries have "obviously one correct answer" or are navigational in nature, search content that could be illegal will rank more highly in response where this is the most relevant to the query, even where an action (such as a penalty) has been applied to the URL that tends to rank it lower overall.[378] [379]

3.65 We accept that relevance is key to search functionality and that downranking cannot prevent a user who is determined to find illegal content from doing so. We have clarified in our measure that the outcome relating to search content being given a lower priority does not necessarily require illegal content to appear lower than other search content in every possible scenario, as outlined above in paragraph 3.40. However, an appropriate moderation action to lower the priority of a search result should result in illegal search content appearing lower when broader search queries are made, or where there are relevant search results available that do not contain illegal content. We also expect the severity of potential harm to users to be considered both in the extent that content is deprioritised, and in whether removing the content from search results would be more appropriate.

3.66 We consider that the updated language in the measure relating to outcome 2 should also assist providers of vertical search services to have greater flexibility. For example, there may be technical actions relevant to some vertical search services operations where a provider could ensure that illegal search content is given a lower priority in results.

3.67 In practice we recognise that providers may prefer to pursue outcome 1 in most cases and take action so that the illegal search content no longer appears in search results. For example, Skyscanner told us that it does not 'downrank' search results. [380]

3.68 In moving to more outcomes-based language, we have designed this measure to ensure services have the flexibility to take appropriate moderation actions that are feasible within the context of their service, whilst still ensuring that providers moderate illegal content in a way that achieves meaningful protections for users.

### Factors relevant to assessing the appropriate moderation action

3.69 We recommend that when designing the aspects of its search moderation function relating to what appropriate moderation action to take (including the extent to which search content is given a lower priority), providers should consider the following factors:

- the prevalence of illegal content hosted at the URL or in the database at which the search content concerned is present;
- the interests of users in receiving any lawful material that would be affected; and
- the severity of potential harm to users if they encounter the content, including whether the content is priority illegal content and the potential harm to children.

[378] See 'Glossary' for definitions.

[379] Google response to November 2023 Consultation, p.39; Google made similar points regarding the use of "downranking" in response to the May 2024 Consultation, p.28.

[380] Skyscanner response to November 2023 Consultation, p.17. Skyscanner made similar points in response to the May 2024 Consultation, p.14.

3.70 These factors are designed to ensure that service providers take action that is appropriate in view of the overall risk presented by the search content. For example, we would expect that URLs containing significant amounts of illegal content, or URLs containing the most severe forms of illegal content, would be actioned such that they no longer appear in search results, rather than just being given a lower priority in overall ranking.

3.71 These factors are intended to enable providers to balance the risks of harm from content against users' rights to freedom of expression. While we recommend that providers have regard to them when designing relevant aspects of their search moderation function, they are not a prescriptive list of factors that are relevant to a consideration of what appropriate moderation action to take. A provider may choose to have regard to the factors in the context of making search moderation decisions about illegal content but would not have to demonstrate that this has been done for each individual decision, provided that the factors have been considered when designing the systems and processes relating to what appropriate moderation action within the provider's search moderation function.

3.72 We have amended the measure to clarify this, which addresses concerns raised by Google and Microsoft outlined at paragraph 3.32. We understand this feedback relates primarily to the potential challenge to moderating at scale were Ofcom to recommend that these factors be considered in the context of every moderation decision.[381]

3.73 We have also clarified that search providers are only expected to assess the prevalence of illegal content hosted at the specific URL or in the database where the search content is present, which we think they can reasonably be expected to assess. This addresses the concern from Google that it would be expected to assess prevalence of all illegal content on a site it does not host[382] (see paragraph 3.32).

3.74 In response to our November 2023 Consultation, stakeholders pointed to additional protection granted to children under the Act.[383] We agree with this and therefore want to explicitly make clear that we are including the risk of harm from illegal content to children as a factor relevant to the severity of illegal content that we expect providers to consider under the Illegal Content Codes. The Act states that search services should be designed and operated in such a way that they provide a higher level of protection for children, including from illegal content.[384] Therefore we are now expressly referencing harm to children as an element of severity, which was listed in our November 2023 Consultation.

### Appeals

3.75 We recognise Google's concern about our proposed measure on appeals in chapter 6 of this Volume: 'Reporting and complaints' and the implications for its entire ranking system, should that be brought within the scope of the appeals that it is required to accept under the Act.[385]

3.76 For the avoidance of doubt, our recommendations about complaints that are appeals will only apply to moderation action that results in that content being removed from search

[381] Google response to November 2023 Consultation, pp. 39-40; Microsoft response to May 2024 Consultation, p.13.

[382] Google response to Nov 2023 consultation, pp.38-39.

[383] ACCO response to November 2023 Consultation, p.4; Hall, J. response to November 2023 Consultation, p.7.

[384] Schedule 5, paragraph 5 (v).

[385] Ofcom/Google meeting, 3 October 2024.

results or being given a lower priority in search results that has been taken on the basis that the content is considered to be illegal content.

3.77 As explained above, we have amended this search measure to give more flexibility for search providers to take action that is consistent with how they operate their service. However, this does not mean that we consider any action taken within a provider's ranking system that impacts the ranking of search content (for example, where this is due to other factors such as authority or accuracy) to be 'search moderation' action as it is referred to in this chapter. Such broader ranking actions would not therefore give rise to a right of appeal for an interested person (i.e. UK-based website or database operator).

3.78 Other action the search provider takes for the purposes of operating its service, that is not taken to comply with the illegal content safety duties, or does not involve making an illegal content judgment, will not fall within the scope of our illegal harm recommendations about complaints that are appeals. For example, if a URL is downranked on the basis that it is poor quality, we would not expect this to fall under scope of the appeals duties. We note that some of these actions may still fall within scope of our recommendations about other types of complaints such as where they involve the use of proactive technology. We set out our recommendations for each type of complaint in chapter 6 of this Volume: 'Reporting and complaints'.

3.79 Where a provider chooses to take alternative measures to meet its safety duties, it will need to consider how these measures impact any other duties it has under the Act, such as complaints duties.

## Benefits and effectiveness

3.80 One of the most important ways in which search providers can reduce the risk of users encountering illegal content of all kinds is effectively implementing their search moderation systems and processes.

3.81 As set out in the 'Search' chapter of the Register of Risks ('Register') a wide range of illegal content can be accessed via search services.

3.82 General search services, through indexing the entire content of the web, carry an underlying risk of harm due to the users' ability to enter search requests and receive search content – which may include any illegal content available online – in the search results. Our evidence base highlights the accessibility of CSAM, terrorism content and extreme pornography, as well as content related to the sale of prohibited items and articles.[386] In the Register, we also set out extensively how exposure to these types of illegal content can cause significant harm.

3.83 This measure will reduce users' exposure to illegal content on general search services, thereby delivering important benefits. While actions that result in content no longer appearing in search results may be more effective at eliminating the risk to users, we recognise that they may not be appropriate in every case. We consider that actions giving search content a lower priority in search results will also contribute to reducing the risk of users encountering illegal content, and of harm being caused by that content (compared to a counterfactual where illegal search content remains easily accessible via search results).

[386] In particular, see Register of Risks chapter titled 'Search' for a full breakdown of the kinds of illegal harm that have been linked with search services.

3.84 Vertical search services have an inherently lower risk given the far narrower scope of content presented to users that comes from pre-determined, often professional, or curated, locations on the web. However, it is still possible that users of vertical search services could encounter illegal search content. In having a search moderation function, users of vertical search services will be protected as necessary in line with the duties under the Act.

3.85 We do not anticipate that all illegal search content will be prevented from appearing in search results in every case. This measure recommends that providers consider the factors outlined in paragraph 3.69 to determine which outcome is appropriate given the prevalence and severity of the content and the rights of UK users to access legal content that might be also available at the same URL or database. In response to stakeholder feedback, we consider deindexing or delisting in all circumstances to be inappropriate.[387] We consider our approach proportionate and beneficial for UK users given their right to access search content without undue interference.

3.86 We consider that this updated measure provides benefits for UK users and practical flexibility for providers. It sets out clear recommendations for how search service providers can fulfil their duties to protect UK users from illegal search content, whilst still affording a level of flexibility for providers to decide how to set up their search moderation functions and what action is most appropriate in different contexts.

## Costs and risks

3.87 The costs of implementation will vary depending on the type and size of service. For providers of small search services that are low-risk and receive few complaints, the costs may be low. For service providers at significant risk of search results that include illegal content, the costs could be considerably higher. The volume of complaints about suspected illegal content may be greater on these services, and the moderation systems and processes required to manage them may need to be more complex and comprehensive. These costs are likely to include both one-off costs of developing a system and ongoing costs of maintaining it. One-off costs for providers that decide to build their own systems internally may include hiring experienced moderation systems designers, developing moderation tools, project management and integration with data analytics/measurement software.

3.88 We consider that the amendments we have made to the measure may result in lower costs to services than our original proposals. This is because the measure is now less specific about the technical action service providers need to take in response to identified illegal content.

3.89 We consider that the costs discussed here reflect the base level of cost which is required to design and operate a search moderation system to review and minimize the risk of harm from encountering illegal search content. We consider that a proportionate approach for large and risker services will also entail costs additional to this, as set out in the additional measures below.

3.90 While the costs described in this section may be significant for some providers, we consider that these measures capture the minimum steps to ensure that they meet their duty to have proportionate systems and processes designed to minimise the risk of online users

[387] [✂]; Protection Group International response to November 2023 Consultation, p.6.

encountering illegal search content. They are also the minimum needed to enable providers to consider complaints about illegal search content appropriately. Incurring these costs is therefore necessary to meet the requirements of the Act.

## Rights impact

3.91 As with content moderation by U2U services (discussed in chapter 2 of this Volume), search moderation is an area in which the steps taken by service providers to comply with their duties under the Act may have a significant impact on the rights of users and those responsible for website or database operators,[388] in particular the right to privacy (Article 8) and freedom of expression (Article 10) under the ECHR.

### Freedom of expression

3.92 As explained in 'Introduction, our duties, and navigating the Statement', as well as in chapter 14 of this Volume: 'Statutory tests', Article 10 of the ECHR sets out the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. It is a qualified right and we must exercise our duties under the Act in a way that does not restrict this right unless we are satisfied that to do so is prescribed by law, pursues a legitimate aim, is proportionate to the legitimate aim, and corresponds to a pressing social need.

3.93 We have carefully considered the impact of this measure on rights to freedom of expression, including the right of search service providers and website operators to impart information to users, and of users to receive information and ideas. We acknowledge that moderation actions taken in line with this measure will impact the ease with which users access illegal content by means of a search service (and in some cases, will prevent access entirely). This impact is potentially significant if a judgement by a provider is incorrect (as in this case, there would not be a substantial public interest in access to the piece of content in question being restricted). However, the duty on search service providers to minimise the risk of users encountering illegal content is a requirement of the Act, and is therefore reflected in this measure. By limiting user exposure to illegal search content, any moderation action taken will be in pursuit of the legitimate aims of preventing crime, protecting health and morals, and protecting the rights of others (particularly victims of crime).

3.94 We consider that the measure is designed in such a way as to minimise the potential impact on freedom of expression, for users, website operators, and providers.

3.95 First, the measure does not recommend that providers take any action against illegal content of which they are not yet aware, nor does it recommend them to restrict access to any content which they do not judge to be illegal content. The measure offers flexibility insofar as it does not recommend removal of illegal content from search results in all circumstances; instead, it recommends that it may be appropriate for the provider to take action that results in the illegal content being given a lower priority in the overall ranking of search results.[389] The factors that we recommend providers have regard to when

[388] In section 227(7) of the Act, this group is referred to as "interested persons" insofar as they are based in the United Kingdom. However, our rights assessment considers the rights of website and database operators more broadly, irrespective of where they are based.

[389] As outlined in paragraph 3.85, we do not consider it appropriate to recommend that providers of search services take action to remove illegal content from search results in all cases which was suggested by two stakeholders, including: [✂]; Protection Group International response to November 2023 Consultation, p.6.

considering appropriate moderation actions, including prevalence, severity of illegal content and impact on content that is not illegal hosted at the same location, are intended to enable providers to balance the overall risk against freedom of expression rights of users and website operators.

3.96 Second, this measure also specifies other Codes measures as safeguards for the free expression rights of users and website/database operators. These safeguards operate in a number of different ways, including ensuring that (where those other measures apply to the service in question) the service provider sets internal content policies and provides training and material to individuals working in moderation. The safeguards would support providers in determining whether detected search content has been accurately identified, and in providing a level of transparency for users and website/database operators about any technology used and how to make a complaint. Additionally, in accordance with the principles of the Act[390] and our duties under the Human Rights Act 1998,[391] we will have regard to the importance of freedom of expression when making any decisions about enforcement in relation to this measure, which acts as an additional safeguard for these rights.

3.97 We therefore consider that to the extent that these measures impact on rights to freedom of expression of users, website/database operators and providers of search services, it is likely to constitute the minimum degree of interference required to secure that service providers fulfil their safety duties about illegal content under the Act.

3.98 There is a potential risk of error in search moderation, for example where a provider makes an incorrect judgement as to the illegality of search content. Impacts on freedom of expression could in principle arise in relation to the most highly protected forms of speech, such as religious expression (which could also affect users' rights to religion or belief under Article 9) or political expression, 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).[392] We recognise that, in certain circumstances, it can be difficult to assess whether such kinds of content should be classified as illegal content, especially when considering whether such content may constitute a hate offence.

3.99 However, the definition of illegal content is statutory. Providers also have incentives to limit the amount of content that is wrongly actioned, both to meet their users' expectations and to avoid the costs of dealing with appeals. In this context, our measures on appeals [from paragraph 6.302 of chapter 6] also act as a safeguard for freedom of expression. We have prepared the Illegal Content Judgements Guidance (ICJG) with careful regard to rights of freedom of expression and encourage service providers to have regard to the ICJG when implementing this measure, to assist with correctly identifying when freedom of expression considerations are particularly relevant to availability of certain content.

3.100 While it is not a requirement of the measure, we acknowledge that a greater degree of interference with these rights could arise if the service provider chose to define the content it restricts more widely than is necessary to comply with the Act. In this case, providers could also be restricting access to certain types of content which is not required under the duties in the Act, and might also not be harmful, or might be less severely harmful.

[390] In particular, see section 1(3)(b).

[391] Section 6.

[392] See the duties set out in sections 17, 18 and 19 of the Act.

However, it remains open to services as a commercial matter (and in the exercise of their own right to freedom of expression) to decide what forms of content to restrict, so long as they comply with the Act. We have no power to prevent them from doing so. If they choose to do so more as a result of the Act, this is the effect of the Act and not our recommendations. However, service providers have incentives to meet their users' expectations in this regard, too.

3.101 Overall, we consider it unlikely that a less restrictive approach to search moderation could be adopted while still enabling service providers to fulfil their illegal content safety duties under the Act. Taking this into account along with the benefits to users and affected persons, we consider that the impact of this measure on the freedom of expression of users, website operators and providers is proportionate.

### Privacy

3.102 As explained in 'Introduction, our duties, and navigating the Statement', Article 8 of the ECHR sets out the right to respect for an individual's private and family life. An 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.

3.103 All search moderation, whether by automated tools or human moderators, will impact on the rights of individuals to privacy and their rights under data protection law (discussed further from paragraphs 3.110 – 3.113 below). The degree of interference with the right to privacy will depend to a degree on the extent to which the nature of the affected content gives rise to a legitimate expectation of privacy.

3.104 Overall, we do not consider that moderation of search content in line with this measure will amount to an undue interference with users' rights to privacy.

3.105 Search content identified as illegal content and actioned through search moderation functions is, by definition, either identified in a way that enables a general search service to have made it available via search results, or made available for publication by a vertical search service under a relevant arrangement with the content provider. This content will not, by its nature, contain information about any users of the service that requires processing in the identification of illegal content or application of an action. In addition, the actions recommended in this measure do not include any action against individual users. However, the process of moderation may include the processing of personal data in the handling of user complaints about suspected illegal content.

3.106 The duty for services to treat illegal content appropriately is a requirement of the Act, and not of this measure, and we are giving services flexibility as to precisely how they implement this and what moderation action they take. We acknowledge that service providers may choose to implement this measure in a way that involves a greater or lesser impact on users' privacy rights. However, as noted above, it remains open to service providers in the exercise of their own rights to freedom of expression to decide what forms of search content to provide access to, as well as what forms of personal data they consider they need to gather to enforce their content polices and give effect to this measure. Providers have this flexibility as long as they comply with the Act and the requirements of data protection legislation (as discussed from paragraph 3.110 onwards).[393] Providers are also required by the Act to have particular regard to users' privacy rights when deciding on and implementing safety measures.[394]

3.107 As identified by C3P, the moderation of some kinds of illegal content can also act to directly protect the rights of victims and survivors, for example those depicted in CSAM.[395] We recognise that certain types of illegal content cause ongoing harm to victims from knowing that the material continues to circulate online and/or from being identified by persons who have viewed that material. Action taken to remove this content from search results protects victims' and survivors' rights under Article 8 ECHR and protects their personal data.[396]

3.108 Where CSAM is identified through the operation of the search moderation function recommended by this measure, providers may in certain circumstances be required (or choose) to report this to a law enforcement authority or to a designated reporting body. Relevantly, section 66 of the Act (which is not yet in force) sets out duties for search service providers to report to the National Crime Agency (NCA) detected CSEA content which is not otherwise reported.[397] Providers may also have additional CSEA reporting duties in other jurisdictions or have voluntary reporting arrangements. Aspects of the Act's reporting duties are to be further defined in regulations made by the Secretary of State.[398] However, a report may include information about individuals responsible for hosting the content or other identifiable individuals (for example, victims or perpetrators who appear in that content or otherwise contribute to the website), which may present an additional risk to the right to privacy. In its consultation response, the ICO queried whether reporting of this nature would have any impact on the rights of users and said that it expected the impact to be limited to third party personal data contained on webpages.[399] We agree that, unlike for U2U services, these duties would not involve the reporting of any user, as the functionalities of a search service do not enable users to upload or share content on the service. We therefore consider that reporting would have an impact on the privacy rights of non-user individuals such as website operators.

3.109 In part, any such interference results from the reporting duties created by the Act or by existing legislation in other jurisdictions. Where CSEA content identified by a search provider is correctly reported in line with the Act, any interference is prescribed by the relevant legislation. In enacting the legislation, Parliament has already made a judgement

[393] Ofcom has given guidance on what information we consider to be reasonably available to service providers for the purposes of making illegal content judgments, in the preparation of which we have had regard to the right to privacy and the principle of data minimisation.

[394] Set out in section 33 of the Act in relation to search services.

[395] C3P response to November 2023 Consultation, p.17.

[396] Review by human moderators of content accurately detected to be CSAM also represents a significant interference with the privacy rights of the victims it depicts. However, that review forms an important part of ensuring that these measures are proportionate and appropriate for service providers to take for the purposes of complying with their illegal content safety duties. We therefore consider that the intrusion into victims' privacy rights is necessary and that no less intrusive approach would be a suitable alternative.

[397] In the case of a non-UK provider of a regulated search service, the duty is limited to "UK-linked" CSEA content: s.66(4).

[398] Section 67 of the Act requires the Secretary of State to make regulations which will set out the information to be included in reports to the NCA and may also require the retention of user-generated content, user data and associated metadata.

[399] ICO response to November 2023 Illegal Harms Consultation, pp.17-18.

that such interference is a proportionate way of securing the relevant public interest objectives. However, we recognise that the risk to the privacy rights of website operators and other non-user individuals will be particularly acute in respect of any content that is incorrectly reported. In that regard, we agree with the ICO that the accuracy principle in data protection law is of particular relevance in the context of reporting CSEA content (as reiterated in paragraph 3.111 below).[400]

### Data protection

3.110 The degree of impact will also depend on the extent of personal data about individuals which may need to be processed to give effect to the applicable search moderation processes. The measure does not specify that service providers should obtain or retain any specific types of personal data about users or other individuals as part of their moderation processes; we give guidance about that separately in our ICJG. We consider that service providers can implement the measure in a way which minimises the amount of personal data which may be processed or retained so that it is no more than is needed to give effect to their moderation processes.

3.111 As outlined in paragraph 2.99 in chapter 2 of this Volume: 'Content moderation', providers should familiarise themselves with the data protection legislation and relevant guidance from the ICO when processing personal data for the purposes of this measure.[401] This means they should apply appropriate safeguards to protect the rights of users, including for example having regard to the need for personal data to be accurate, in line with the accuracy principle in data protection law. We note that this may be particularly relevant in the context of the provider reporting any CSEA content identified to the NCA or other law enforcement agency as described in paragraphs 3.108 – 3.109 above. Providers may also use third parties to carry out search moderation on their behalf and ICO guidance is clear that providers should ensure that individuals' (such as reporting users) rights to privacy are fully protected when a third party has access to their personal data.[402]

3.112 In line with feedback from the ICO in response to our measures recommending content moderation by U2U services (discussed in the 'Content moderation' chapter),[403] we have also updated this search moderation measure to include specific references to the privacy safeguards provided by other measures which apply to certain providers operating a search moderation function. This clarifies the protections afforded to individuals by the Codes and how this measure seeks to minimise the impact on individuals' privacy rights.

3.113 Overall, we consider that (assuming service providers also comply with applicable data protection legislation requirements and guidance) the privacy impact of this measure as a result of a provider's search moderation decisions and processes, above and beyond the requirements of the Act, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their safety duties about illegal content under the Act. Taking this, and the benefits to users into consideration, we consider that it is proportionate.

[400] The ICO said that the accuracy principle in data protection law "requires that [service providers] take all reasonable steps to ensure that the personal data they process is not incorrect or misleading as to any matter of fact".
[401] ICO, UK GDPR guidance and resources and Online safety and data protection [accessed 4 November 2024]
[402] Further information on the requirements for contracts between data controllers and processors can be found at Contracts and liabilities between controllers and processors.
[403] ICO response to November 2023 Illegal Harms Consultation, p.12.

### Who this measure applies to

3.114 This measure is recommended to apply to all search service providers. We explain why we have taken this position in the Conclusion section.

## Conclusion

3.115 We have concluded that it would be proportionate to include a measure in the Codes stipulating that providers of search services should review and assess search content that they have reason to suspect may be illegal content, and take appropriate moderation action that results in illegal content no longer appearing in search results or being given a lower priority in the overall ranking of search results. We consider that search services would be highly unlikely to be able to comply with their duties under the Act without implementing such a measure.

3.116 Moreover, ensuring that content which providers have determined to be illegal either does not appear in search results – or, at a minimum, is given lower priority in such results – will reduce users' exposure to such content, thereby protecting users and delivering important benefits. Whilst we have not been able to quantify the costs, we do not consider them likely to be disproportionate given the measure is in effect necessary to comply with the law and given that we have allowed search providers significant flexibility and discretion in how they do it.

3.117 As it is highly unlikely that a provider could comply with its duties under the Act without following this measure, we have applied it to all search services (noting our clarification regarding expectations on downstream search services as outlined in the 'Our approach to developing Codes measures' chapter).

3.118 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C1. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

# Measure on internal search content policies

3.119 In our November 2023 Consultation, we proposed that internal search moderation policies are set having regard to the findings of the provider's risk assessment and any evidence of emerging harms on the service. We proposed this measure should apply to all providers of large general search services and all providers of multi-risk search services.

3.120 In our proposed amendments to the Codes (which we consulted on in May 2024 alongside the Children's Safety measures), we altered the reference to "emerging harm" and instead recommended that services should have processes in place to update these policies in response to any evidence of new and increasing illegal harm on the service.[404]

3.121 We considered that search moderation policies help to secure more accurate, consistent, and timely decision-making.

[404] We did this to clarify the meaning of "emerging harm" in the measure, and to clarify action providers should take according to this.

## Summary of stakeholder feedback[405]

3.122 Besides those stakeholders who expressed broader support for the full package of search moderation measures in this chapter (outlined in paragraph 3.14), we received no responses specifically addressing this measure. Furthermore, we received no responses to our proposed amendments in the May 2024 Consultation.

## Our decision

3.123 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation, including the subsequent amendment consulted on alongside the May 2024 Consultation. We have made a small number of minor clarificatory changes:

- We have replaced the references to “deindexing” and “downranking” with “taking appropriate moderation action”, in line with the changes made to measure ICS C1 in response to stakeholder feedback.[406]
- Our measure now says that in setting and recording internal content policies to take appropriate moderation action, that providers should have processes in place for updating these policies in response to evidence of new and increasing illegal harm on the service (as tracked in accordance with the measure outlined in paragraph 5.158 onwards in Volume 1: chapter 5: ‘Governance and accountability’). This is to clarify that we are not recommending providers update their internal content policies every time they receive evidence of new and increasing illegal harm on their services, but that they have processes in place to do so where appropriate.

3.124 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C2. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

## Our reasoning

### How this measure works

3.125 Search moderation typically relies on general rules, or search moderation policies, that apply (in principle) to all search content that is made available on or via a search service. With regard to general search services, policies are generally applied to individual URLs or domains, and this is often done at scale by providers of larger services.[407]

3.126 Similar to the content moderation policies of U2U services, we understand search moderation policies to exist in two forms:

- External policies are publicly available documents aimed at users of the service which provide an overview of a service provider’s rules about what content is allowed and what content is restricted. It is a requirement of the Act that providers publish a publicly

[405] Note this list in not exhaustive, and further responses can be found in Annex 1.

[406] We have amended this wording to reflect the new approach taken in measure ICS C1. As such, references to “deindexing” and “downranking” have been replaced by “taking appropriate moderation action” in line with the outcomes identified there. We also refer to content that is restricted (rather than prohibited) in publicly available statements, in further accordance with changes to Measure ICS C1.

[407] Google, no date. Content policies for Google Search. [accessed 25 November 2024]; Ofcom, 2019. Use of AI in Content Moderation. [accessed 25 November 2024]; Google, no date. Our approach to information quality and content moderation. [accessed 25 November 2024]; Ofcom, 2023. Content moderation in user-to-user online services: An overview of processes and challenges.

available statement that includes provisions specifying how individuals are to be protected from illegal content (for example, through moderation), and our recommendations on this are in chapter 10 of this Volume: ‘Terms of service and publicly available statements’.

- Internal policies are usually more detailed versions of external policies and may set out rules or standards for staff involved in search moderation. Once internal policies are set, they can be used as a guide for enforcement by search moderators and other relevant teams, as well as to assist in identifying potential breaches by designers of automated systems.[408]

3.127 Search moderation policies can help ensure more accurate and consistent decision-making, particularly in organisations where moderation is carried out by a large team.

3.128 We recommend that search service providers set and record internal policies setting out rules, standards, and guidelines around what search content will be subject to moderation action, as well as explaining how policies should be operationalised and enforced. These should be drafted to enable search service providers to take appropriate moderation action in line with our measure ICS C1.

3.129 In setting and recording internal content policies, the measure specifies that providers should have regard to their illegal content risk assessment and have processes in place for updating these policies in response to evidence of new and increasing illegal harm on their services (as set out in Volume 1: chapter 5: ‘Governance and accountability’).

3.130 In line with our chapter on U2U content moderation, we have amended the wording of this part of the measure to clarify that we are not recommending that providers should update their internal content policies every time they receive evidence of a new and increasing illegal harm. Rather, we are recommending that they should have processes in place to be able to do this where appropriate. Providers may be able to take other actions to protect users in response to evidence of new and increasing illegal harms on their services that does not require them to update their internal content policies, and the measure is drafted to account for this.

## Benefits and effectiveness

3.131 This measure will contribute to tackling the harms that may result from insufficient moderation systems and processes (as outlined from paragraph 3.80 onwards).

3.132 Providers of large general search services may face significant challenges due to both the volume and diverse nature of the content they need to moderate. This can make it difficult to ensure consistency in decision-making across large moderation teams. Service providers identifying risks of multiple kinds of illegal harms are also likely to face these challenges.

3.133 These challenges raise questions about how providers should:

- prioritise content for review;
- achieve consistency, quality, and timeliness of decision-making; and

[408] Khoury College at Northeastern University, no date. Content Moderation Techniques. [accessed 25 November 2024]; Trust and Safety Professional Association, no date. Policy Development. [accessed 25 November 2024]; Google, no date. Content policies for Google Search. [accessed 25 November 2024].

- plan their deployment of moderation resourcing to secure that users are appropriately protected.

3.134 We consider content policies to be a necessary step to ensure effective moderation on search services. In chapter 2 of this Volume: 'Content moderation', we noted that stakeholders expressed a broad consensus that setting internal content policies is a necessary first step to establishing an effective content moderation system for some U2U services. We also noted that where services are larger or higher risk (requiring providers to moderate large volumes of diverse content), clear content moderation policies are vital for ensuring consistency, accuracy, and timeliness of decision-making. We consider the same to be true for search service providers and their search moderation functions. Setting content policies therefore plays a central role in preventing users from being exposed to illegal content, thereby delivering important benefits.

3.135 We see significant benefits in recommending that service providers have regard to their illegal content risk assessments when setting and recording their policies. Risk assessments will provide evidence about the challenges faced by service providers' moderation functions. It is reasonable to infer that such data would enable providers to make higher quality decisions about what to put in their internal content policies for search moderation (tailored to the needs of their specific services). Further, the Act requires that service providers carry out a risk assessment when they make 'significant changes' to their services, providers may update their internal policies in response to any increases in risks of certain harms created by these changes. We do not consider it is necessary to make more specific recommendations about what should be included in internal content policies to achieve such benefits.[409]

3.136 We also think there are likely to be significant benefits in recommending that service providers have processes in place for updating their internal content policies in response to any evidence of new or increasing illegal harm on the service. Where policies are kept updated, search moderation decisions are more likely to be based on the most recent data. They will therefore likely be of higher quality and enable providers to better protect users from harms through search moderation.

### Costs and risks

3.137 Service providers that do not currently have an internal content policy will incur the costs of developing them. Some service providers may choose to use external experts, which could increase costs. Approving new policies may also take up senior management's time, which would add to the upfront costs. Since our November 2023 Consultation, we have further analysed these costs for the purposes of our May 2024 Consultation. We estimated that the cost to providers of smaller search services of implementing the equivalent measure in the Children's Safety Codes could be in the region of £3,000 to £7,000.[410] While this cost estimate relates to developing an internal content policy relating to content harmful to children, we expect that the costs of developing such a policy relating to illegal harms could

[409] This includes reference to Glitch's response to our proposal in the November 2023 Illegal Harms Consultation that we outline services carry gender-specific policies. Source: Glitch response to November 2023 Illegal Harms Consultation, p.6. We address this in more detail in chapter 2 of this Volume: 'Content moderation'.

[410] This is assuming a service required three weeks of time across professional occupations (legal/regulatory staff) and four hours of senior management time to develop an internal search moderation policy. This is based on our wage estimate assumptions as set out in Annex 5 of this document.

be similar for many smaller providers. This is because the development process, staff involvement, and time required is likely to be similar. For both types of harm, costs are likely to differ between providers depending on the type and number of harms present on the service.

3.138 Providers of large services may require more complex content policies, as the way in which harm can materialise is likely to be more varied on such services and the governance requirements needed to implement them are also likely to be more complex. These factors may increase costs due to the increased amount of time required to design more complex policies. These costs could reach the tens of thousands or more.[411] There may also be some small ongoing costs to ensure these policies remain up to date over time (for example, to take into account new and increasing illegal harms).

3.139 We understand large general search services already have such policies in place, which means that in practice this measure might only impose costs for providers if those services relating to ensuring their internal policies are sufficient to meet their duties under the Act.

3.140 These costs are mitigated by the flexibility of the measure, as we have set out high-level recommendations that give providers flexibility over how they choose to implement them.

## Rights impact

### Freedom of expression

3.141 We consider that this measure has the potential to impact on the rights to freedom of expression for the reasons set out in relation to the measure on taking appropriate moderation action in relation to illegal search content (ICS C1). This is because the internal content policy would inform the provider’s moderation decisions made according to that measure.

3.142 In addition to those impacts, we consider that this measure has the potential to interfere with the right to freedom of expression of users and website/database operators if the internal content policies define the content in scope of these policies more widely than is necessary to comply with the Act. However, nothing in this measure requires or encourages providers to do this. As a matter of their own right to freedom of expression, providers are entitled to decide what content they want to present to users in response to search requests, and how to present it, so long as they protect UK users from the types of harmful content (including illegal content) regulated by the Act.[412]

3.143 We consider there may also be positive impacts on users’ right to freedom of expression from providers implementing this measure. Internal moderation policies can set out a level of detail that may not be practical to do in external facing policies, providing moderators with greater clarity on the type of search content that is illegal content and priority illegal content, resulting in a higher degree of content being identified appropriately. Where services are likely to be dealing with large volumes of search content, the process of considering these matters in advance and preparing a policy would tend to improve internal scrutiny and improve the consistency of decisions in a way which we consider would also tend to protect users’ rights to freedom of expression.

[411] These cost estimates do not change the approach on which we consulted in our November 2023 Consultation, but add further detail to support our position.

[412] Section 33(2) of the Act.

3.144 We therefore consider that the impact of this measure on the right to freedom of expression, above and beyond the requirements of the Act, is likely to constitute the minimum degree of interference required to secure that service providers fulfil their safety duties about illegal harms under the Act. Taking this, and the benefits to users into consideration, we consider that it is therefore proportionate.

### Privacy and data protection

3.145 We do not expect this measure to result in any additional interference with users' rights to privacy under Article 8 or their rights under data protection law beyond those identified above in relation to the measure on taking appropriate moderation (ICS C1). Where the internal content policy describes or defines content, this is likely to be by reference to search content that is generally available to be presented to users by operation of the underlying search engine and would not, by its nature, contain information in relation to which a user is unlikely to have a reasonable expectation of privacy.

3.146 Providers processing personal data in the implementation of this measure will need to comply with applicable data protection legislation[413] and will separately be required to ensure that privacy duties under the Act are met.[414] Having a set of policies in place will also encourage consistency and predictability in search moderation, which will help to secure that any processing of personal information is appropriate. We therefore consider that (assuming service providers comply with applicable data protection laws) this measure is likely to constitute the minimum degree of interference required to secure that service providers fulfil their illegal content safety duties under the Act. Taking this, and the benefits to users into consideration, we consider that it is therefore proportionate.

## Who this measure applies to

3.147 For the reasons set out in paragraph 3.132 and as discussed in our November 2023 Consultation, it is likely to be difficult for providers of large general search services and providers of any other search services identifying risks of multiple harms to carry out effective moderation without internal content policies. Given the importance of effective search moderation to ensuring that users are protected from search content that is illegal content, we expect the benefits to users of applying this measure to such services to be substantial.

3.148 We understand that providers of the largest general search services already have content policies in place and are therefore unlikely to incur substantial new costs as a result of applying this measure. We expect that costs are unlikely to be significant for smaller multi-risk search services. We therefore consider the measure to be proportionate for large general search services and those presenting risk of multiple harms.

3.149 We consider that automatically applying the measure to providers of large vertical search services would have a materially smaller benefit for UK users. As set out in our Register chapter title 'Search', our analysis suggests the risks of illegal harms on vertical search services are relatively low, as such services draw results via an API (or equivalent technical means) from pre-determined websites that may contain professional or curated content (such as particular products or services), rather than indexing sites from across the clear

[413] In determining what this requires of them, providers should have regard to any relevant guidance from the ICO. For example, see: ICO, UK GDPR guidance and resources and Online safety and data protection [accessed 4 November 2024].

[414] Section 33 of the Act sets out privacy duties for search services.

web. Given the lower risks, the volume and complexity of complaints received by providers of vertical search services about potentially illegal content is likely to be materially smaller than for general search services and the benefits of having internal content policies would be lower. However, if it were the case that any vertical search service was multi-risk, then we consider the measure to be proportionate, as with other multi-risk services.

3.150 We are therefore recommending this measure for all providers of large general search services and all providers of multi-risk search services.

3.151 As discussed in 'Our approach to developing Codes measures', we may consult in 2025 on extending this measure to some or all single-risk services.

## Conclusion

3.152 For the reasons set out above, we have concluded that applying this measure to the search providers in question would be beneficial and proportionate.

3.153 Therefore, we have decided to leave the measure largely unchanged from the measure proposed in our November 2023 Consultation, alongside the subsequent amendment in the May 2024 Consultation, except for two new minor amendments. We have replaced references to "deindexing" and "downranking" with "taking appropriate moderation action", in line with the changes made to measure ICS C1. We have also outlined that providers should "have processes in place for updating policies in response to evidence of new and increasing illegal harm on the service", rather than update their internal policies every time in response to such evidence.

3.154 This measure applies to large general search services and other search services which are multi-risk.

3.155 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C2. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

# Measure on performance targets for search moderation functions

3.156 In the November 2023 Consultation, we proposed that providers should set performance targets for their search moderation function and measure whether they are achieving them. We proposed that these should include the time that illegal content remains on the service before it is deindexed or downranked, and the accuracy of decision making. We described that when setting targets, providers should balance the desirability of deindexing or downranking illegal content swiftly against the need to make accurate moderation decisions.

3.157 We proposed this measure apply to all providers of large general search services and all providers of multi-risk search services.

3.158 We considered it important that service providers are clear about the search moderation outcomes they are trying to achieve and are measuring whether they are achieving them. We believed this would enable them to configure their systems appropriately to meet these goals and be able to optimise the operation of these systems.

## Summary of stakeholder feedback[415]

3.159 Our analysis of stakeholder responses identified the following areas of concern. Specifically:

- concerns about the practical applicability of time targets; and
- concerns about unintended incentives created by performance targets.

3.160 We address these concerns throughout the 'Our reasoning' section. We address areas additional stakeholder responses in Annex 1.

### Concerns about the practical applicability of time targets

3.161 Google raised concerns that the measure as initially set out, particularly the target relating to the length of time that illegal content remains on a service, lacks clarity.[416] Having considered this feedback, we have clarified our recommendations under section 'How this measure works' (paragraphs 3.167 – 3.171).

### Concerns about unintended incentives created by performance targets

3.162 Several respondents raised concerns about the risk of adverse impacts on freedom of expression, if providers are incentivised to moderate content at pace by performance targets relating to time.[417]

3.163 We address these concerns in the 'Benefits and effectiveness' section (paragraphs 3.182 – 3.188) and the 'Freedom of expression' sub-section (paragraph 3.194).

## Our decision

3.164 We have decided to recommend the measure broadly as we proposed in the November 2023 Consultation. We have made a number of clarificatory changes in response to the feedback set out in the previous section.

- We have made minor changes to align with amendments to our measure on taking appropriate action in relation to illegal search content. As such, references to "deindexing" and "downranking" have been replaced by "taking appropriate moderation action".
- We have decided to amend this measure to clarify the period over which time taken to action illegal content should be measured. It should begin when a provider first has reason to suspect a given piece of search content may be illegal and should end when appropriate moderation action is taken in line with our measure ICS C1. This process is explained in further detail under 'How this measure works'.
- In accordance with the measure on performance targets in chapter 2 of this Volume: 'Content Moderation', providers also now have the option to set performance targets for search content that is illegal content or an **illegal content proxy**, aligned with the choice they make, as described in paragraphs 3.47 – 3.49 above. We have done this by

[415] Note this list in not exhaustive, and further responses can be found in Annex 1.

[416] Google response to November 2023 Consultation, p.41.

[417] Big Brother Watch response to November 2023 Consultation, pp.4-5; Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) response to November 2023 Illegal Harms Consultation, p.8; Protection Group International response to November 2023 Consultation, p.6; Zevo Health response to November 2023 Illegal Harms Consultation, p.8.

making an express reference to **appropriate moderation action** being taken in line with measure ICU C1 (which covers both illegal content and illegal content proxy).

- Finally, the measure also now says that providers should balance the **need** to take appropriate moderation action swiftly with the **importance** of making accurate moderation decisions. These words replace the term **desirability** which we proposed in the November 2023 Consultation. This is to clarify our expectation that providers must balance speed and accuracy to set appropriate performance targets for their services.

3.165 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C3. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

# Our reasoning

## How this measure works

3.166 In line with our general approach to search moderation, we are not prescribing exactly what performance targets should be. However, at a minimum, we recommend that performance targets cover at least:

a) the time from which providers first have reason to suspect a given piece of search content may be illegal to when appropriate moderation action has been taken; and

b) the accuracy of decisions made regarding whether search content is illegal content or an illegal content proxy.

### Targets for the time taken to action illegal content

3.167 We note Google's concern that it would be difficult to identify the point in time at which content is 'on the service' (for example, where content may be indexed but never served in response to a query) in line with the original drafting of this measure.[418] In particular, we acknowledge that there is no onus on search providers to proactively assess websites for illegal content generally (with the exception of the CSAM URLs measure outlined in chapter 5 of this Volume: 'Automated search moderation').

3.168 To clarify when time targets should apply, the measure now recommends that providers should set targets for the time it takes for them to review and take appropriate moderation action in respect of identified illegal search content.

3.169 In its response to the November 2023 Consultation, Google noted that "it is difficult to specify to what extent a result is downranked with precise attribution as to why that downranking occurred".[419] We are clarifying that the time threshold ends when appropriate moderation action is taken in line with measure ICS C1, not on the conclusion of a specific technical action which we recognise may only occur, in the case of action taken that results in content being given a lower priority in the overall ranking of search results, in response to individual user requests.

3.170 This amendment to the framing of the time target is consistent with the measure's aim of encouraging timely moderation to improve protections for users. It also makes the measure more consistent with measure ICS C1 (in which we shift focus to moderation outcomes rather than prescribing the precise technical actions of 'deindex' and 'downrank').

[418] Google response to November 2023 Consultation, p.36.

[419] Google response to November 2023 Consultation, p.41.

3.171 We are not aware of any performance targets currently used by providers of large general search services regarding the average time taken to act on illegal content. However, Google noted that its reporting mechanisms are designed to allow users to provide information for Google Search to quickly assess and act where necessary.[420]

### Targets for the accuracy of decision making

3.172 We also recommend that the performance targets include targets for the accuracy of decision-making. In chapter 2 of this Volume: 'Content moderation', we give the example of some U2U providers doing this by tracking the rate of appeals as a measure of the accuracy of decisions that are taken. We know that Microsoft Bing tracks accuracy metrics to monitor moderation effectiveness.[421]

3.173 We do not consider it necessary to be any more prescriptive in defining accuracy of decision-making. We consider that providers are best placed to set appropriate performance targets for accuracy based on what is most suitable for their services, including the extent to which they use information about complaints and appeals to inform such targets.

3.174 We acknowledge that a focus only on speed-based or time-based performance targets may result in poor quality decisions. Our measure aims to mitigate this risk by not specifying time targets for services and by recommending that services set accuracy targets in addition to time-based performance targets. This will ensure that a focus on speed of decision-making is balanced against a focus on accuracy. Services will be made aware of any decline in accuracy rates, meaning that they will be in a better position to respond to underperformance.

### Balancing the need to take appropriate moderation action swiftly with the importance of accuracy of decision making

3.175 We recognise that services will need to determine the appropriate balance between targets for time and accuracy to help ensure the quality of search moderation practices, and we note that the importance of this balance has been highlighted by some stakeholders at paragraphs 2.181-184 in chapter 2 this Volume: 'Content moderation'. We consider that the appropriate balance for each service will be subject to the specific risks and needs of that service.

3.176 In setting its targets, the provider should balance the need to take appropriate moderation action swiftly in relation to illegal content or illegal content proxy, with the importance of making accurate moderation decisions.

3.177 We have replaced the term **desirability** with the words **need** and **importance** in this measure to clarify our expectation that speed and accuracy are not only desirable, but are essential components of an effective search moderation system. Providers should set their performance targets in a way that pursues both speed and accuracy of moderation and does not solely pursue one of these factors to the detriment of the other. As explained in the 'Content moderation' chapter at paragraph 2.210, we consider that the tension between these two factors is a beneficial feature of this measure and incentivises providers to strike a balance between these factors, making their performance targets more effective at protecting users on their service.

[420] Google response to 2022 Illegal Harms Call for Evidence, p.21.
[421] Microsoft Bing, 2023. Bing EU Digital Services Act Transparency Report. [accessed 25 November 2024].

3.178 We would expect a provider to be able to justify why it has set the performance targets it has, including why the targets in question are reasonable and how it has balanced the need for speed and accuracy when making this decision.

3.179 We do not consider there to be any tension between providers setting time and accuracy targets for search moderation and also having other wider performance metrics for which they set targets. We would welcome providers wanting to design a range of targets related to user safety that are appropriate to the risks on their services and decision-making processes that go beyond the types of performance targets listed in this measure.

### How this measure fits with other Search Moderation measures

3.180 In accordance with the measure on performance targets in our 'Content moderation' chapter, we have made an amendment to our measure to clarify that providers now have the option to set performance targets for illegal content or illegal content proxy. This better reflects our position in ICS C1 that a provider may either make an illegal content judgement or assess search content against its publicly available statement.

3.181 We do not recommend the outcomes that performance targets should achieve. However, we specify in the measure on resourcing (paragraph 3.257) that providers should resource their search moderation functions to give effect to their performance targets. We consider that these two measures together will ensure that search moderation functions are sufficiently resourced to meet performance targets.

## Benefits and effectiveness

3.182 Zevo Health argued that there could be a knock-on effect on user protection, as welfare issues and staff turnover due to increased pressure on moderators could impact how quickly and accurately moderators will be able to remove illegal content.[422] Big Brother Watch added that setting targets for the time taken to remove content will create pressure on companies to remove content at pace, which has implications for freedom of expression.[423]

3.183 Similarly, Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) said performance targets rely on the prerogative of platforms to make determinations on the illegality of content, that should only be retained for CSAM content through automated hash systems. It also said that providers should not be making determinations on the illegality of the content, and that the time in which illegal content remains online is not a good metric as this could create incentives for a rapid removal, deindexing, or downranking of any content deemed suspicious.[424]

3.184 We agree that a disproportionate focus on speed of content removal could lead to pressure on moderators, resulting in poorer quality decisions. This could have potential adverse effects on user protection from illegal search content and on the right to freedom of expression of users, website owners, and service providers. However, as described in paragraphs 3.172 – 3.173, we are allowing providers flexibility regarding how to structure their targets and have explicitly recommended that they should balance the need for both speed and accuracy of decision-making.

[422] Zevo Health response to November 2023 Consultation, p.8.

[423] Big Brother Watch response to November 2023 Consultation, pp.4-5.

[424] CELE response to November 2023 Consultation, p.8.

3.185 Where providers explicitly set targets and measure performance against them, they are more likely to be able to optimise the design of their moderation functions to achieve the goals underlying the targets than they would be if they did not set targets. This is because we consider that where a provider sets and measures its performance against performance targets, as long as it is incentivised to meet the targets, it is more likely to achieve the things that the targets relate to. For example, we consider that all else being equal, a provider with an overall aspiration of swiftly actioning illegal content would be more likely to do so if it set clear and explicit targets for timeliness of moderation action than if it did not.

3.186 As we have explained at 3.134, timeliness and accuracy of decision making are essential to an effective search moderation function. While it is important that services moderate illegal content quickly to protect users, it is equally important that the decisions they make are accurate (as inaccurate decisions could either result in illegal content remaining accessible when it should not be or result in legal content being over-moderated).

3.187 Providers that balance the need to take appropriate moderation action in relation to illegal content with the importance of making accurate moderation decisions are more likely to strike an appropriate balance between speed and accuracy of moderation decisions. This will benefit users both by making it more likely that illegal content is actioned promptly and by making it more likely that search services make accurate moderation decisions.

3.188 Users are only protected if moderation decisions are made in a timely and accurate way. We consider that the safety duties outlined in the Act imply a need for search providers to act swiftly in identifying and acting on illegal search content where proportionate.[425] While we do not consider it appropriate to prescribe precise performance targets, we maintain that there are important benefits to providers setting (at minimum) both time-based and accuracy-based targets for their search moderation teams.

## Costs and risks

3.189 Service providers will incur one-off costs in designing and setting up suitable performance metrics and targets. This may involve one-off system changes (for example, to measure relevant information). To assess the accuracy of search moderation decisions, providers are likely to need to take a sample of these decisions and review them, which could result in significant ongoing costs. There would also be other ongoing costs, including data storage costs.

3.190 Since our November 2023 Consultation, we have further analysed these costs for the purposes of our May 2024 Consultation. Regarding initial implementation costs, we set out that we expect the costs of implementing the equivalent measure in the Children's Safety Codes through the creation of a simple bespoke system to be approximately £8,000 to £16,000.[426] This would be the case where accuracy was estimated based solely on the outcome of user appeals. While this cost estimate relates to developing performance metrics and targets relating to content harmful to children, we expect that the costs of developing such a system relating to illegal harms could be similar for many providers. This

[425] The safety duty for search services, unlike the takedown duty for U2U services, does not include the word "swiftly". However, we consider that user protection implies a need to act swiftly where it is proportionate to do so, so services should at least turn their minds to the need to act swiftly.

[426] This is based on our assumption that this would require around 30 days of software engineering time and our cost assumptions set out in Annex 5 of this document.

is because the development process, staff involvement, and time required is likely to be similar. Alternatively, providers might opt to license a third-party system at a relatively low cost (such solutions are available from around £50 per month for each staff user). However, the cost of designing and implementing more complex systems, tracking a more extensive set of metrics, and carrying out proactive quality assurance of report accuracy would introduce complexity which may significantly impact on the cost. As such, depending on the service design and/or volume of reports, we estimate that initial implementation costs in such cases could run from the tens to hundreds of thousands of pounds.[427]

3.191 We have not quantified the ongoing costs of assessing the accuracy of search moderation decisions as they would depend in part on the complexity of the targets that service providers set, and the volume of content assessed.

## Rights impact

3.192 This measure recommends that providers of in-scope services should set performance targets as part of their internal content policies recommended by Measure ICS C2. This measure should therefore be seen as part of a package of measures relating to search moderation for illegal content, including our measures on taking appropriate action in relation to illegal search content (ICS C1) and on internal search content policies (ICS C2), for which we have assessed the rights impacts at paragraphs 3.91 – 3.113 and 3.141 – 3.146.

### Freedom of expression

3.193 We have not identified any specific additional adverse impacts from this measure on the right to freedom of expression of users, website or database operators, or service providers (beyond those identified in relation to the measure on taking appropriate action in relation to illegal content ICS C1).

3.194 In response to concerns raised by stakeholders, we recognise that the risks to freedom of expression associated with search moderation may be increased by the addition of performance targets (particularly those relating to speed) as these can cause moderators to take decisions quickly, increasing the risk of error.[428] However, as explained in paragraph 3.184, this risk is mitigated by the flexibility of the measure, which recommends that time targets are set at a level which strikes a balance with accuracy. We consider that this reduces the risks to freedom of expression that may arise with more prescriptive time targets for the removal of illegal content, such as those set by the Network Enforcement Act 2017 in Germany (NetzDG).[429] Additionally, our measure includes the recommendation that services also set performance targets for accuracy, which should mean that both speed and accuracy are considered by services, resulting in greater transparency and consistency in search moderation systems. We consider this potentially would have a positive impact on the rights to freedom of expression of users and website or database operators. As outlined in paragraph 3.173, it will be for service providers to ensure that the targets they set are appropriate to mitigate this risk.

[427] These cost estimates do not change the approach on which we consulted in our November 2023 Consultation but add further detail to support our position.

[428] Big Brother Watch response to November 2023 Consultation, pp.4-5; CELE response to November 2023 Consultation, p.8; Protection Group International response to November 2023 Consultation, p.6.

[429] Big Brother Watch response to November 2023 Illegal Harms Consultation, p.5.

3.195 The flexibility of the measure also means that providers have scope to set different performance targets for different circumstances – for example, where there is nuance involved with search moderation decisions – to ensure that accuracy is balanced appropriately against speed of decision-making.

3.196 We recognise that there are a range of factors affecting the likelihood of error, such as issues with automated technology, turnover of moderation staff, time pressure, and moderator experience. We consider that the recommendation for service providers to effectively track their performance against targets (particularly those relating to accuracy) acts as a safeguard for the right to freedom of expression of users, website or database operators and search providers.

3.197 We therefore consider that any interference with the right to freedom of expression of users and website or database operators would be mitigated by the flexibility of the measure, which recommends that time targets are set at a level which strikes a balance with accuracy, and is, as such, relatively limited and proportionate.

### Privacy and data protection

3.198 We consider the privacy and data protection impacts of this measure to be inextricably linked.

3.199 We note the risk that setting speed-based performance targets can lead to a focus on speed rather than accuracy. This could interfere with the privacy rights of individuals since it may lead to the creation of inaccurate personal data. Therefore, we have designed this measure so that services will need to balance the speed of decisions made with the degree of accuracy, which we consider will mitigate the risk of undue interference with an individual's rights.

3.200 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.[430] This will be particularly important when making decisions about CSEA content, where an incorrect decision could lead to individuals being reported. We consider the measure to be compatible with data protection requirements. We do not consider that it would be appropriate for us to duplicate data protection requirements on the face of the measure in the Codes.

3.201 We therefore consider that any interference to users' rights to privacy arising from this measure would be proportionate.

## Who this measure applies to

3.202 As discussed in our November 2023 Consultation, for general search service providers that need to make many moderation decisions, we consider there to be important benefits from setting performance targets for search moderation functions and tracking whether these are met. As outlined in section 'Benefits and effectiveness', we maintain that providers following this measure are more likely to operate effective search moderation systems.

3.203 While the overall costs of this measure are somewhat unclear, we consider the benefits likely to be sufficiently important to justify applying the measure to providers of large

[430] In determining what this requires of them, providers should have regard to any relevant guidance from the ICO. For example, see: ICO, UK GDPR guidance and resources and Online safety and data protection [accessed 4 November 2024].

general search services and search services with risks of multiple harms, given the important role effective search moderation plays in protecting users from harm.

3.204 The lower risks associated with vertical search services make the benefits of this measure likely to be smaller when applied to such services. For this reason, we do not consider it appropriate to apply measure ICS C3 to large vertical search services purely based on their size. That said, we maintain that this measure would be proportionate if a vertical search service were to be assessed as having risks of multiple harms, due to the higher volume of content that would likely require moderation in relation to these services.

3.205 We therefore consider that measure ICS C3 is appropriate for all providers of large general search services and all providers of multi-risk search services.

3.206 As discussed in 'Our approach to developing Codes measures', we may consult in 2025 on extending this measure to some or all single-risk services.

### Conclusion

3.207 Setting and measuring performance targets will materially improve the likelihood that the search moderation function achieves the objectives underlying these targets. It is of foundational importance that search moderation functions make timely and accurate decisions. Therefore, notwithstanding the uncertainty over the precise costs of the measure, we conclude that this is a proportionate intervention. This conclusion is reinforced by the fact that Consultation responses did not contain any clear and compelling evidence that the measure in question would impose disproportionate costs.

3.208 As set out in the preceding section, this measure applies to large general search services and search services which are multi-risk.

3.209 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C3. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

## Measure on a policy for the prioritisation of content for review

3.210 In the November 2023 Consultation, we proposed that a provider should prepare and apply a policy to determine the prioritisation of content to review. In setting the policy, the provider should have regard at least to: how frequently search requests for the search content are made, the severity of potential harm to users if they encounter the content, and the likelihood that the search content is illegal content. We proposed this measure apply to all providers of large general search services and all providers of multi-risk search services.

3.211 We considered that the adoption of a prioritisation framework, and considering these factors in setting this framework, would likely result in providers making high-quality decisions about what search content to prioritise for review, resulting in a material reduction in harm to users.

## Summary of stakeholder feedback[431]

3.212 Our analysis of stakeholder responses identified the following areas of concern, which are summarised in this section:

- Concerns about the practical application of the measure; and
- the inclusion of harm to children as a consideration.

3.213 We address these concerns throughout the 'Our reasoning' section.

### Concerns about the practical application of the measure

3.214 Google said that the measure was too prescriptive in setting out specific factors to be taken into account when prioritising content for review, arguing that it would not always be necessary or appropriate to consider these for every harm or for every service, and that doing so could slow down the moderation process.[432]

3.215 These comments are addressed in paragraph 3.221 under the 'How this measure works' section.

### Inclusion of harm to children as a consideration

3.216 Two civil society stakeholders argued that children should be included in prioritisation frameworks, through either depicted age of a person or general harms to children being factors for prioritisation.[433] This feedback was provided in response to our counterpart measure for U2U services, but we consider it equally relevant to search moderation.

3.217 We address these concerns in paragraph 3.225 under the 'How this measure works' section.

## Our decision

3.218 We have decided to recommend the measure broadly as we proposed in the November 2023 Consultation. We have made two minor clarificatory changes to the measure:

- We have clarified that by severity, we mean the **severity of potential harm to UK users if they encounter illegal search content on the service**.
- In response to the feedback at 3.215, the measure now explicitly states that **potential harm to children** is an aspect to be considered as part of the severity of content.
- We have clarified that by 'how frequently search requests for the search content are made', we mean how frequently the search content is returned in response to search requests.

3.219 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C4. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

[431] Note this list is not exhaustive, and further responses can be found in Annex 1.
[432] Google response to November 2023 Consultation, p.40.
[433] ACCO response to November 2023 Consultation, p.4; Refuge response to November 2023 Illegal Harms Consultation, p.12.

# Our reasoning

## How this measure works

3.220 We recommend that providers of large general search services or multi-risk search services should prepare and apply a policy on prioritising search content for review. In setting the policy, the provider should have regard to at least the following factors:

- how frequently the search content is returned in response to search requests;
- the severity of potential harm to UK users if they encounter the search content on the service (including whether the content is priority illegal content, the risk assessment of the service and the potential harm to children); and
- the likelihood that the search content is illegal content, including whether it has been reported by a trusted flagger.

3.221 Given the immense amount of content in the indexes they maintain, providers of large general search services may need to deal with huge volumes of reports regarding URLs containing potentially illegal content. Service providers identifying risks of multiple harms may also have to respond to high volumes of reports of harmful content. This means providers face difficult decisions about what search content to prioritise for review. Having a policy that takes into account relevant factors should result in high-quality decisions about what search content to prioritise for review.

3.222 It is not our expectation that providers consider each individual factor in the context of every prioritisation decision that is made where it is not relevant to do so. Rather, the measure makes clear that the factors should be considered in setting the policy itself. This addresses feedback from Google that our proposed measure would be onerous to implement in practice were the expectation to be that the prioritisation factors be considered in every instance of moderation.[434]

### How frequently the search content is returned in response to search requests

3.223 Terms that are searched more often and by a greater number of users may indicate a higher risk of harm to users where the content returned is found to be illegal. This factor is intended to consider where there are queries frequently made returning results containing illegal content. This should mean illegal search content affecting the greatest number of UK users is prioritised.

3.224 There is evidence of an existing practice among service providers of considering this frequency in prioritisation frameworks. Google Search considers factors associated with the level of harm, including the volume and frequency of search requests, when prioritising content for review.[435]

3.225 Recognising the importance of search request frequency, it is important to consider this alongside other factors listed at 3.219. Solely prioritising content that frequently returned in response to search queries for may mean other serious harms are missed. For example, websites designed to help criminals commit serious offences may not be commonly searched for but could cause very serious harm.

[434] Google response to November 2023 Consultation, p.40.

[435] Google, 2023, Fraud research note to Ofcom.

### Severity of potential harm to UK users

3.226 Providers should consider the severity of potential harm to UK users if they encounter the search content on the service when setting up their prioritisation policies. This includes whether content is suspected to be priority illegal content, the risk assessment of the service, and the potential to harm children.

3.227 This factor ensures the most harmful illegal search content is prioritised (to the extent that it is possible for providers to identify this before reviewing it). We recommend providers should have regard to the severity of potential harm when designing their prioritisation policy, but do not consider they must necessarily have regard to it in every case. We recognise that providers will differ in what they will be able to understand about search content before they have looked at it. For example, some providers ask complainants to categorise their complaints. Such providers would be able to write a policy which, for example, prioritised a complaint about CSEA over a complaint about copyright infringement. Some providers may also be able to correlate signals surrounding a complaint or a piece of content with a likelihood of severe harm. It would be appropriate for those providers to apply this knowledge in their prioritisation processes. If a provider chose to use automated content detection to identify content for review, and the automated tools it used could detect aspects of content which are associated with very severe harm, it would also be appropriate to take this into account in its prioritisation processes.

3.228 As outlined in paragraph 2.292 in the 'Content moderation' chapter, we know that several U2U service providers already consider the severity of potential harm when prioritising content for review, and that some harms are considered more severe than others. We are aware that some search services may already prioritise based on severity of harm.

3.229 We recommend that providers consider whether the search content is suspected to be priority illegal content as an indicator of severity, because 'severity' is one of the three factors the UK Government used to determine its list of priority illegal offences.[436]

3.230 We also recommend that providers consider the findings of their risk assessments regarding severity of potential harm when setting prioritisation policies, to prioritise types of content that present a particularly severe risk of harm on their service (but may sit outside priority offences).

3.231 We accept that search providers may choose to assess search content against their publicly available statement, rather than by conducting an illegal content judgement (as outlined in paragraph 3.47 in the measure on taking appropriate moderation action in relation to illegal search content). However, search content of a kind identified in publicly available statements as being subject to moderation action may include content that is both legal and illegal. We recommend that search providers should have regard to severity of potential harm to UK users if they encounter the illegal search content when prioritising content for review, including specifically whether the content is suspected to be priority illegal content, which poses a greater risk of harm. Providers may include other aspects of severity of potential harm to UK users in their prioritisation decisions (in addition to the factors listed in paragraph 3.219) based on what is appropriate for their search service. For example, we do not consider that providers should interpret our explanation of severity to

[436] Department for Digital, Culture, Media & Sport, Home Office, The Rt Hon Nadine Dorries MP, and The Rt Hon Priti Patel MP, 2022. Online safety law to be strengthened to stamp out illegal content. [accessed 25 November 2024].

mean that priority illegal content should always be prioritised above the categories of content harmful to children as defined in the Act.

3.232 We note the suggestion from two civil society stakeholders that providers should also consider harms to children when setting their prioritisation policies.[437] As outlined in paragraph 2.307 in chapter 2 of this Volume: 'Content moderation', we agree that the potential of content to be harmful to children is an important indicator of the severity of content and note that providers of search services are obliged to provide a higher standard of protection for children than for adults within the online safety objectives listed in the Act.[438] We have therefore amended this measure to clarify that potential harm to children is an aspect of the severity of potential harm to which providers should have regard when setting their search moderation prioritisation policy.

### The likelihood that the search content is illegal content, including whether it has been reported by a trusted flagger

3.233 We recommend that providers have regard to the likelihood that content is illegal, including whether it has been reported by a trusted flagger, in setting their prioritisation policies.

3.234 There are numerous ways that providers may have reason to suspect that search content is illegal. These may include user complaints, proactive detection, or reports from a trusted flagger.[439]

3.235 Complaints and reports by users are likely to be a valuable way for providers to find out about illegal content, particularly for those not making extensive use of proactive detection methodologies. However, in chapter 2 of this Volume: 'Content moderation', we note that users are not always correct in identifying breaches of U2U services' content policies. We consider the same is likely to be true of search services. As such, we consider that another indicator of the likelihood that content is illegal is whether it has been flagged by a trusted flagger.

3.236 The Dedicated Reporting Channel ('DRC') measure (chapter 6 of this Volume: 'Reporting and complaints') sets out instances in which we recommend that providers make a reporting mechanism available to named trusted flaggers in relation to fraud. As noted in chapter 6, the DRC measure does not prevent the use of the reporting channel for the reporting of other illegal content or intelligence by other trusted flaggers who are assessed by the provider to have sufficient expertise.

3.237 In setting up its prioritisation policy, a provider should have regard to the likelihood that search content is illegal, and one factor in determining whether it is illegal will be that it has been reported by a trusted flagger. We have only recommended that providers establish trusted flagger arrangements with entities which we consider can be expected to flag content correctly, and we do not recommend that providers establish relationships with trusted flaggers unless they are asked to do so by the trusted flagger concerned. Our measure also leaves it open to providers to have a policy for prioritising content for review which is not based on this factor, so long as in setting their policy they have considered whether and how to prioritise flags from trusted flaggers. For example, in chapter 5 of this

[437] ACCO response to November 2023 Consultation, p.4; Refuge response to November 2023 Consultation, p.12.

[438] See Schedule 4 paragraph 4(a)(vi) and, more generally, section 1(3)(b)(i) of the Act.

[439] The broad description of who a trusted flagger is can be found in chapter 2 of this Volume: 'Content Moderation'.

Volume: 'Automated search moderation', we consider certain kinds of automated technology which are associated with a high likelihood that content they identify is illegal. The likelihood that the content is illegal is self-evidently relevant to whether further review is needed and how quickly it should take place.

### Other factors

3.238 We deal with additional factors suggested by stakeholders in chapter 2 of this Volume: 'Content moderation' at paragraph 2.270. We are not proposing that services have regard to a specific harm as part of their prioritisation process, due to the risk of this giving rise to unintended consequences.

3.239 We consider that this measure gives providers the flexibility to incorporate our recommended prioritisation factors into their prioritisation frameworks as they see fit (provided these factors are considered when setting up policies for the prioritisation of content for review).

## Benefits and effectiveness

3.240 We maintain that adopting a prioritisation framework for review of content can result in a material reduction in harm compared to a chronological approach in which search providers simply reviewed complaints in order of receipt.

3.241 Our approach delivers significant benefits. The prioritisation framework we set out:

- ensures that illegal content with the potential to cause harm to larger audiences is prioritised, thereby protecting more users;
- addresses the most severe content, minimising the harm to users from priority illegal offences and other illegal harms that are identified through the risk assessment, as well as protecting children as part of that consideration; and
- considers evidence that search content is likely to be illegal, thereby increasing the efficiency with which this content is actioned, and users protected.

3.242 In summary, effective prioritisation will ensure the most severe illegal content that is reaching large numbers of users is reviewed quickly, minimising the risk of users encountering that content.

## Costs and risks

3.243 The creation of a prioritisation policy will not in itself have an impact on the overall amount of search content that providers of search services need to review. However, there will be costs incurred in designing and applying the prioritisation policy and these will largely be one-off in nature. Designing the prioritisation policy may take several weeks of fulltime work and involve legal, regulatory, and ICT staff, as well as experts in harms and online safety. Agreeing new policies may also require input from senior management, which would add to the upfront costs. Applying the prioritisation policy may require system changes, such as ensuring that the frequency of queries is taken into account when content is reviewed by content moderators and ensuring that content from trusted flaggers is suitably prioritised. There may be material one-off costs in making these changes.

3.244 Since our November 2023 Consultation, we have further analysed these costs for the purposes of our May 2024 Consultation. We anticipate that designing and setting up a relatively simple prioritisation framework could cost approximately £4,000 to £7,000.[440] While this cost estimate relates to developing a prioritisation system relating to content harmful to children, we expect that the costs of developing such a system relating to illegal harms could be similar for many smaller providers as the development process, staff involvement, and time required is likely to be similar. However, for a larger and more complex service – with a multitude of different metrics that can indicate virality, severity, and suspected type of content – costs could be substantially higher than this, potentially reaching tens of thousands or more. This reflects the more complex design requirements and set-up costs for tools such as ticketing systems or systems that automate what content is reviewed next.[441]

3.245 There may also be some smaller ongoing costs incurred ensuring that the prioritisation policy is reflected in system design and in reviewing it when appropriate. We consider all these costs will be mitigated by the measure not specifying exactly how service providers should prioritise content, giving them some flexibility how they achieve this.

3.246 As the amount of content reviewed may not change, it is not clear that implementing this measure will impose other material ongoing search moderation costs on providers compared to a counterfactual in which they simply reviewed complaints chronologically. Indeed, having a clear prioritisation framework may help them deploy their resources more efficiently.

## Rights impact

3.247 This measure should be seen as part of a package of measures relating to search moderation for illegal content, including Measures ICS C1 and ICS C2, for which we have assessed the rights impacts at 3.91 – 3.113 and 3.141 – 3.146.

### Freedom of expression

3.248 We do not consider that setting and applying a policy for the prioritisation of search content for review, in itself, has any specific adverse impacts on the right to freedom of expression of users, website or database operators, or service providers. It may have a positive impact on the right to freedom of expression, as the recommendations of the measure mean that harm would be a factor in service providers' decision-making and that users would be able to more safely engage with communities and content online.

### Privacy and data protection

3.249 We consider that setting and applying a prioritisation policy would only have additional impacts on users' privacy or personal data rights beyond those already considered, to the extent that it involved a further use of private information or processing of personal data by the provider concerned. However, any such extra processing would need to be carried out in compliance with applicable privacy and data protection laws and so we do not consider that it would be disproportionate.

## Who this measure applies to

3.250 As outlined in our November 2023 Consultation, we expect that providers of large general search services and search services identifying risks of multiple harms will have a large

[440] This is assuming it would require three weeks FTE from professional occupations (legal, regulatory, ICT) and one day from senior management and our salary assumptions as set out in Annex 5 of this document.

[441] These cost estimates do not change the approach on which we consulted in our November 2023 Consultation but add further detail to support our position.

quantity of potentially illegal content to review. We consider that the benefits of applying this measure to providers of such search services is likely to be particularly great because they are likely to receive large volumes of reports of potentially illegal content. Therefore, there will be particular benefits of such services having a policy for the prioritisation of search content for review – both to providers in the form of better decision-making and more efficient use of resources, and consequently to users through minimisation of their risk of encountering illegal content. By providing recommendations for prioritising the review of such content, this measure aims to ultimately reduce users' experiences of harm from illegal content.

3.251 As set out in paragraph 3.105, we maintain that the lower risks associated with vertical search services make the benefits of this measure likely to be smaller when applied to such services. For this reason, we do not consider it appropriate to apply measure ICS C4 to large vertical search services purely based on their size. That said, we maintain that this measure would be proportionate if a vertical search service were assessed as having risks of multiple harms due to the higher volume of content requiring assessment.

3.252 We therefore consider that measure ICS C4 is appropriate for all providers of large general search services and all providers of multi-risk search services.

3.253 As discussed in 'Our approach to developing Codes measures', we may consult in 2025 on extending this measure to some or all single-risk services.

## Conclusion

3.254 Our analysis shows that the measure we are recommending is likely to deliver material benefits by improving the quality of decisions about what content to prioritise for review, and in doing so, better protect users. Whilst the measure will have some costs, these are relatively modest. Given the importance of ensuring that providers of services that receive large volumes of reports of potentially illegal content prioritise them appropriately, we consider these costs to be proportionate. Therefore, we have decided to leave the measure largely unchanged from the measure we proposed in our November 2023 Consultation, except for the addition of "potential harm to children" as an indicator of the severity of potential harm to UK users in the measure and the clarifications in respect of severity and frequency. We have also clarified in paragraph 3.219 how we expect this measure to operate, noting that the factors outlined are relevant to the setting of the prioritisation policy but need not be considered in the context of each individual prioritisation decision.

3.255 This measure therefore recommends that service providers should prepare and apply a policy for the prioritisation of search content for review. In setting this policy, we recommend that providers have regard at least to how frequently the content is returned in response to search requests, the severity of potential harm to users if they encounter the content, and the likelihood that the search content is illegal (including where it is flagged by a trusted flagger).

3.256 This measure applies to large general search services and search services which are multi-risk.

3.257 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C4. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

# Measure on resourcing search moderation functions

3.258 In our November 2023 Consultation, we proposed that providers should ensure that their search moderation teams are resourced to give effect to their internal content policies and performance targets. In doing so, we set out that providers should have regard to at least the propensity for external events that lead to a significant increase in demand for search moderation, and the particular needs of its United Kingdom user base. We proposed this measure should apply to all providers of large general search services and all providers of multi-risk search services.

3.259 We considered that this measure would help service providers review potentially illegal content faster and make more accurate decisions about whether to remove it. The success of our recommendation on performance targets also relies on providers resourcing their search moderation functions to meet those targets.

## Summary of stakeholder feedback[442]

3.260 Alongside responses to our November 2023 Consultation supporting the requirement for search moderation functions, we received one response that specifically supported our proposal to recommend that providers adequately resource their search moderation functions to prevent their users from encountering illegal content.[443]

## Our decision

3.261 We have decided to proceed with the measure as proposed in our November 2023 Consultation. The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C5. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

## Our reasoning

### How this measure works

3.262 We recommend that service providers resource their search moderation functions sufficiently to meet their internal content policies and performance targets set in line with measures ICS C2 and ICS C3 as outlined earlier in this chapter. In line with our approach to search moderation, we do not consider it appropriate to specify in detail how providers should resource their search moderation functions. However, we recommend that providers should have regard to at least the following factors when deciding how to resource their search moderation functions:

- the propensity for external events to lead to a significant increase in demand for search moderation on the service; and[444]

[442] Note this list in not exhaustive, and further responses can be found in Annex 1.

[443] [✂].

[444] Information obtained from service providers' risk assessments, tracking evidence of new kinds of illegal content and other relevant sources of information could be used to understand where and when some such occurrences might happen.

- the particular needs of their UK user base as identified in their risk assessment, in relation to languages.[445]

### The propensity for external events to lead to a significant increase in demand for search moderation

3.263 In instances where systems may need to deal with sudden significant increases in illegal search content or unexpected harm events, redeploying resources to do so may draw resources away from another part of the system. It is beneficial for search service providers to consider the potential for sudden significant increases in problematic (and potentially illegal) content when determining how to resource their search moderation functions.

### Resourcing for the particular language needs of UK user bases

3.264 The provider should consider the particular language needs of its UK user base as identified in its risk assessment. This means that if a large proportion of the UK user base is likely to use the search service in certain languages, then the search moderation function should be equipped to moderate search content in those languages accordingly.

3.265 We expect that providers should be prepared to adapt to changing prevalence in search languages across their UK users.

## Benefits and effectiveness

3.266 Providers with large volumes of content or many different types of content to review are unlikely to be able to keep users safe using ad hoc methods and without specialist resourcing. Providers are likely to need specific resources to handle complaints, and may need to adjust their overall resources and how they use them based on what is happening on their service.

3.267 We therefore consider that adequate resourcing of search moderation functions will result in providers making more accurate and timely decisions about what appropriate moderation action to take in relation to the illegal search content identified. We would expect this to result in a material reduction in harm to users and deliver significant benefits.

3.268 Responses to our 2023 Protection of Children Call for Evidence stress the importance of adequately resourcing moderation functions of U2U and search services. The Center for Countering Digital Hate (CCDH) highlighted the need for providers to improve their content moderation functions through substantial resourcing and dedicated human moderators to deliver greater protections for children online.[446] We consider this to be true in terms of protecting all users from illegal harms.

3.269 This aligns with the evidence discussed in the Register of Risks chapter titled ‘Governance, systems and processes’ which concerns how limited resourcing, time pressures, and large or fluctuating volumes of content requiring moderation can contribute to increased risk.

3.270 We also consider that setting objectives in relation to time and accuracy of a search moderation function will not protect users unless the provider also has sufficient resources

[445] This measure is not prescriptive about the specific language expertise or resources providers should use. This is because the risk of harm in a particular language will likely differ from service to service based on a number of factors, including userbase. Service providers should use their most recent risk assessment (which should include analysis of their userbase) in reaching this judgement. It should be noted that the Act is concerned with protecting users of services in the UK, meaning any recommendation would be in relation to languages used or viewed by users of services in the UK.

[446] Center for Countering Digital Hate (CCDH) response to 2023 Protection of Children Call for Evidence, p.10.

and deploys them effectively as set out in section 3.265. We therefore conclude that adequately resourcing search moderation functions to meet performance targets will bring significant benefits to users.

### The propensity for external events to lead to a significant increase in demand for search moderation

3.271 Evidence suggests that providers need to build flexibility into their search moderation functions to be effective. In response to the 2022 Illegal Harms Call for Evidence, Business for Social Responsibility stressed the importance of providers "investing in the capability to scale-up/scale-down on short notice to respond to crisis events that can result in sudden spikes in illegal content".[447] For example, search service providers may experience sudden significant increases in complaints about search content at times when there is significant public concern about a particular issue. Users may be at a heightened risk of encountering illegal content if providers fail to take proportionate steps to plan for this. We consider it will therefore be directly beneficial to users for providers to be adequately prepared for these increases.

### Resourcing for the particular language needs of UK user bases

3.272 We know users in the UK use search services in multiple languages.[448]

3.273 Harm is likely to be reduced where service providers ensure their search moderation processes include the language skills needed to moderate potentially illegal content which could be encountered by and harm these users. Considering the specific needs of the UK user base through the risk assessment will ensure that resourcing for language needs is most relevant and therefore most effective.

3.274 In paragraphs 2.363 – 2.365 of chapter 2 of this Volume: 'Content moderation', we set out evidence that U2U services moderate multilingual content and highlight the importance of providers being able to deal with different languages and understand cultural context. This is supported by the stakeholder responses outlined in paragraph 2.363 of that chapter, as well as studies which highlight the need for moderators to understand the cultural context of the content they moderate, including its language.[449]

3.275 By recommending that language proficiency be factored into the resourcing of search moderation functions, we consider that this measure will result in materially better protections for users of the service than if those skills were not available. For the same reasons given the 'Content moderation' chapter, we do not consider it would be appropriate at this time for our recommendation on language resourcing to be more prescriptive.

3.276 Based on our analysis, we conclude that general resourcing of different languages (such as moderators with language expertise or automated systems that work in the required language) would enable providers to take appropriate moderation action more accurately regarding search content that is suspected to be illegal.

[447] Business for Social Responsibility (BSR) response to 2022 Illegal Harms Call for Evidence, p.8.

[448] Vox, 2015. In which language do you Google? Tracking 135 languages in 9 cities since 2004. [accessed 25 November 2024].

[449] British and Irish Law, Education and Technology Association (BILETA) response to November 2023 Illegal Harms Consultation, p.10; Electronic Frontier Foundation response to November 2023 Illegal Harms Consultation, p.10; Open Rights Group response to November 2023 Illegal Harms Consultation, p.2.

## Costs and risks

3.277 The costs of resourcing a provider's search moderation function to give effect to its internal content policies and meet performance targets are likely to be substantial and ongoing. It will tend to be higher in the case of search services that provide access to a greater volume of content, such as when more webpages are included in the index – which is likely to be the case for providers of larger general search services.[450]

3.278 Costs are likely to vary depending on the type of detection and review processes used. Automated moderation processes (such as machine learning solutions for artificial intelligence) require both one-off infrastructure investment and time input from various ICT professionals. Ongoing costs may also be incurred from system updates and licensing. If search moderation involves human moderators, resourcing costs will primarily depend on how many moderators are needed. To be effective, human moderators may require specific training (see Measure 6). They may also need an ICT support team. Service providers may decide to offer mental health support and other wellbeing benefits to their search moderators and other staff working on search moderation, which would add to costs.[451] Some service providers may require a separate review process for more complex illegal search content cases, which may require legal input.[452]

3.279 While these costs will be significant for some providers, this measure does not recommend specific resourcing targets and it will be for providers to determine what they need to do to meet their duties as required by the Act.

## Rights impact

3.280 This measure should be seen as part of a package of measures relating to search moderation for illegal content, including measures ICS C1 and ICS C2, for which we have assessed the rights impacts at 3.91 – 3.113 and 3.141 – 3.146. We do not consider that the measure will have any additional negative impact on the rights of users, website or database operators, or service providers. Appropriately resourcing search moderation is likely to have positive impacts on, and safeguard, those rights, because mistakes are less likely and because the result should be that users feel safer using the service.

## Who this measure applies to

3.281 We have decided to focus this measure on providers of large search services as well as providers of multi-risk services because the size of or risks associated with such services make it especially important that they resource their search moderation functions appropriately. We have decided not to apply it to providers of vertical search services that are not multi-risk given the lower risks associated with these services.

[450] Based on submissions from these parties, Google's index contains around [500-600 billion] pages and Microsoft's index contains around [100-200 billion] pages". Source: CMA, 2020. Online platforms and digital advertising market study final report, pp. 89-90. [accessed 25 November 2024].

[451] Where search moderation is performed by employees of a provider, the provider will need to consider its duty of care to these employees and which the provider may consider involves offering such support and benefits.

[452] "Our legal removals team, comprising trained experts, reviews the report and determines whether to remove the content in accordance with applicable laws." Source: Google response to 2022 Call for Evidence, p.23.

3.282 As discussed in ‘Our approach to developing Codes measures’, we may consult in 2025 on extending this measure to some or all single-risk services.

## Conclusion

3.283 The analysis set out in this section, coupled with the evidence we have received in consultation responses and other engagement with stakeholders, shows that adequately resourcing search moderation functions is both important and beneficial. Where service providers do not resource their search moderation functions adequately, it is unlikely that they will be able to achieve good safety outcomes for users. Although the costs of resourcing search moderation functions have the potential to be very significant, we therefore consider that this measure is proportionate and necessary. This view is reinforced by the fact that no consultation responses argued against the measure under consideration.

3.284 This measure applies to large general search services and search services which are multi-risk.

3.285 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C5 This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

# Measure on the provision of training and materials to individuals working in moderation (non-volunteers)

3.286 In our November 2023 Consultation, we proposed that people working in search moderation should receive training and materials to enable them to moderate search content in line with measures ICS C1 and ICS C2. In doing so, we recommended the provider should:

a) have regard to at least the risk assessment of the service and information pertaining to the tracking of signals of emerging illegal harm; and
b) where the provider identifies a gap in moderators’ understanding of a specific kind of illegal harm, that it gives training and materials to remedy this.

3.287 We proposed this measure apply to all providers of large general search services and all providers of multi-risk search services.

3.288 In our proposed amendments to the Illegal Content Codes consulted on alongside the May 2024 Consultation, we amended the reference to “signals of emerging illegal harm” to clarify that providers should consider “evidence of new and increasing illegal harm on the service (as tracked in accordance with Recommendation A3.13)”. The title of the measure was also amended to clarify that it applies only to paid moderators.[453]

3.289 For service providers subject to measure ICS C2, we consider it unlikely that moderators will be able to implement internal content policies effectively without adequate and appropriate training and materials.

[453] Ofcom, 2024. Amendments to Illegal Content Codes of Practice for user-to user services and search services.

## Summary of stakeholder feedback[454]

3.290 While a range of stakeholders agreed with the need for the package of search moderation measures proposed overall, some respondents suggested amending this measure to recommend that service providers carry out more training on specific harms and user rights. We summarise this feedback in the following sub-section.

3.291 We received no responses with respect to the amendments made alongside our May 2024 Consultation.

### What training should include

3.292 Samaritans noted particularly the expectation to address gaps in paid moderator's understanding related to suicide and self-harm.[455] Glitch felt that training for moderation teams needed to adequately address the intersectional nature of online harm as well as cultural sensitivity.[456]

3.293 We address these responses in paragraph 3.308 in the section entitled 'How this measure works' below.

## Our decision

3.294 We have decided to recommend the measure broadly as we proposed in the November 2023 Consultation (along with the amendments proposed in the May 2024 Consultation). We have made a minor change to clarify that we intend this measure to relate to training of all people (who are non-volunteers) working on search moderation.

3.295 Our measure therefore now says that providers must provide training to **individuals** working in search moderation that enables them to **"fulfil their role in moderating content"**, instead of "to moderate search content". It also now says that this recommendation is "**including in relation to"** instead of "in accordance with" the measures on taking appropriate moderation action and their internal content policy. These amendments are to acknowledge that individuals working in search moderation could have roles in the wider ecosystem of moderation and may not be directly moderating search content themselves.

3.296 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C6. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

## Our reasoning

### How this measure works

3.297 Our view is that providers are best placed to determine what is appropriate for their services in terms of the content of training and materials. However, we recommend that in ensuring that individuals working in search moderation (who are not volunteers) receive training and materials that enable them to fulfil their role in moderating content, a provider should ensure that:

[454] Note this list in not exhaustive, and further responses can be found in Annex 1.
[455] Samaritans response to November 2023 Illegal Harms Consultation, p.3.
[456] Glitch response to November 2023 Illegal Harms Consultation, p.7.

- it has regard to at least the illegal content risk assessment of the service and evidence of new and increasing illegal harm on the service (in accordance with the measure outlined in Volume 1: chapter 5: 'Governance and accountability' regarding the tracking of emerging illegal harm); and
- where the provider identifies a gap in the understanding of individuals working in search moderation in relation to a specific kind of illegal harm, it gives training and materials to remedy this.

3.298 We are aware that some providers of larger services train their moderators and other relevant members of staff to identify and action illegal content. For example, Microsoft Bing ensures human reviewers receive extensive training on its policies.[457]

3.299 A number of civil society organisations stressed the importance of training moderation staff in their responses to the 2022 Illegal Harms Call for Evidence.[458] The importance of training is also supported by broader academic research.[459]

### Individuals working in search moderation functions

3.300 We recommend that individuals working in a provider's search moderation function should receive training and materials that enable them to fulfil their role, including in accordance with our measure on identifying and taking appropriate action in relation to illegal search content, and their internal content policy. This training need not be provided to any volunteers working in search moderation (although we are currently unaware of these being used by search service providers).

3.301 We expect the paid staff working in search moderation would mostly be moderators employed or contracted by providers, though it could include those who are involved in the wider search moderation ecosystem, as described in further detail chapter 2 of this Volume: 'Content Moderation'. We are clarifying this further in our amendment to the Code wording.

### Materials

3.302 Providers should supply materials to people working in search moderation to enable them to fulfil their functions as set out in the other measures in this chapter. This includes sufficient training to understand and operationalise internal content policies, as well as to take appropriate moderation action where appropriate.

3.303 Specific materials provided to people working in moderation may include definitions and explanations around specific parts of internal policies, enforcement guidelines, examples, and visuals of the review interface (the tool or interface moderation staff will use to carry out their job). What is provided may vary depending on a number of factors, including the

[457] Microsoft, 2023. Bing EU Digital Services Act Report. [accessed 25 November 2024].

[458] 5Rights Foundation response to 2022 Illegal Harms Call for Evidence, p.10; Glitch response to the 2022 Illegal Harms Call for Evidence, p.5; Global Partners Digital response to the 2022 Illegal Harms Call for Evidence, p.11; Carnegie UK response to 2022 Illegal Harms Call for Evidence, p.10; CCDH response to the 2022 Illegal Harms Call for Evidence, p.7; NSPCC response to the 2022 Illegal Harms Call for Evidence, p.10; Refuge response to the 2022 Illegal Harms Call for Evidence, p.8; Samaritans response to the 2022 Illegal Harms Call for Evidence, p.7.

[459] Ofcom, 2019. Use of AI in Content Moderation. [accessed 25 November 2024]; Alan Turing Institute, 2021. Understanding online hate: VSP Regulation and the broader context. [accessed 25 November 2024]; Brennan Center for Justice, 2021. Double Standards in Social Media Content Moderation. [accessed 25 November 2024].

type of service, the type of content being moderated, and the local laws and regulations of the region where the service operates.

### Frequency of training

3.304 We note that some respondents raised the benefits of regularly updating the training of individuals working in moderation, as assessed in paragraph 2.431 in chapter 2 of this Volume: ‘Content Moderation’.[460] While we acknowledge the benefits of regularly updating training, we do not propose to prescribe how often training materials should be updated or how often training should be delivered as we consider providers to be best placed to determine what is appropriate to respond to the specific needs and risks of their service and staff functions.

3.305 There is no set best practice on how often training or supporting materials should be refreshed, and it may depend on several factors, including a person’s role and performance, the risks of illegal harm a service faces, and the extent to which such risks vary over time. Therefore, we do not consider that it would be appropriate to specify in Codes how often materials should be revised, or training should be repeated.

3.306 However, a provider which failed to refresh training and materials following any major changes to policies, or processes relating to the moderation of suspected illegal content or proxy content, would not be enabling its moderators to meet the requirements of the other search moderation measures and would therefore not be compliant with the Codes.

### Having regard to risk assessments and any new and increasing harm on a service

3.307 Providers should have regard to their risk assessments to identify areas that they may need to focus training on. As those working in search moderation should be focussed on enforcing their internal content policies, it makes logical sense for training to be informed by the results of the most recent illegal content risk assessment.

3.308 For example, if a service is high risk for illegal suicide and self-harm content, providers should ensure their staff are appropriately trained in the subject matter so that they are able to take appropriate moderation action.

3.309 We consider this link to risk assessment to be more appropriate than being prescriptive about the specific harms that people are trained in. This ensures that providers are focusing their training on areas of highest risk to their user base. We consider that this addresses points raised by stakeholders outlining that services should provide training in specific harm areas.[461]

3.310 In Volume 1, chapter 5: ‘Governance and accountability’, we recommend that providers should track signals of new and increasing illegal harm. This is a crucial source of information that should be used to inform moderator training alongside the regard to risk assessments.

[460] C3P response to November 2023 Consultation, p.16; Global Partners Digital response to November 2023 Illegal Harms Consultation, p.13; Meta response to November 2023 Illegal Harms Consultation, p.26.

[461] Samaritans response to November 2023 Consultation, p.3; Glitch response to November 2023 Consultation, p.7.

### Remedying gaps in the understanding of individuals working in search moderation in relation to specific kinds of illegal harm through training

3.311 Providers should ensure that where they identify a gap in the understanding of individuals working in search moderation in relation to a specific kind of illegal harm, they give training and materials to remedy this.

## Benefits and effectiveness

3.312 Providing training and materials to paid individuals working in search moderation will contribute to tackling the harms that may result from insufficient moderation systems and processes (as outlined in paragraphs 3.82 – 3.83).

3.313 In line with measure ICS C1, a core element of search moderation involves assessing whether a particular item of search content is illegal content (whether by making an illegal content judgement or by assessing the search content against the types of content identified in a provider's publicly available statement) and considering what moderation action is appropriate to minimise the risk of users encountering it. It follows that moderators will need to know how to conduct these assessments in order to carry out this work.

3.314 For service providers subject to measure ICS C2, we consider it unlikely that people working in moderation will be able to implement internal content policies without training and additional materials.

3.315 Based on the information in section 'How this measure works', we maintain that training individuals involved in moderation and providing them with relevant materials will be beneficial for identifying and minimising the risk of users encountering illegal content. Those that have been trained on how to identify and action content in accordance with measure ICS C1 are more likely than untrained individuals to be equipped with the knowledge and skills to identify when action needs to be taken against search content. They are therefore likely to make better search moderation decisions than untrained individuals, thus resulting in users of search service being exposed to less illegal content than would otherwise be the case.

### Having regard to risk assessments and any new and increasing harm on a service

3.316 There will be significant benefits if search service providers have regard to their risk assessments and evidence of any new and increasing harm on the service when determining what training to provide. Where service providers ensure their moderation teams are adequately trained in the harms for which they are at high risk, these teams will be better able to protect users of the search service from the most relevant illegal content.

3.317 Where moderation teams are adequately trained in new harms which are increasing in prominence on the search service, they will be better able to reduce users' exposure to these harms.

### Remedying gaps in the understanding of individuals working in search moderation in relation to specific illegal harms through training

3.318 Where individuals working in search moderation have been trained on how to identify and action content in accordance with measure ICS C1 and measure ICS C2, they are more likely than untrained individuals to be equipped with the knowledge and skills to identify when action needs to be taken against search content. The training should enable them to fulfil the role that they have in moderating content, whatever that might be.

3.319 There may be instances where individuals working in moderation do not have sufficient understanding of specific harms to enable them to effectively minimise the risk of users encountering illegal content. Harms-specific training and materials may be helpful in equipping individuals to identify and action search content that is illegal content due to the unique, complex, novel, or serious nature of a given harm, or because certain harm or harms may be particularly prevalent on a service and so require more in-depth understanding. If training and materials are provided to those working in search moderation where a service provider has identified a gap in its understanding of a specific harm, this should improve outcomes for users.

3.320 Where teams working in moderation lack knowledge of specific harms, this gives rise to a risk of errors being made in moderation decisions. This is because there is a greater chance that content is miscategorised when those working in moderation do not fully understand the harm area. This aspect of the measure is designed to address this.

## Costs and risks

3.321 The main factors driving the cost of the training would be the number of individuals to be trained and the duration of the training. Our analysis of this is the same as that outlined in chapter 2 of this Volume: 'Content moderation' (see paragraphs 2.444 to 2.451). In summary, we estimate the cost of providing training to be between £3,000 and £18,000 for a new search moderator and between £5,000 and £28,000 for a new software engineer.[462] If search moderation staff are based in countries with lower labour costs than the UK, the lower end of the assumed wage range may overstate the costs. Costs may also vary depending on whether the training is given by in-house staff or by an external provider.

3.322 In addition to the costs of training new staff working in search moderation and software engineers, there will also be some ongoing costs for refresher training and training in new harms emerging on services. We expect that the annual costs of these would be lower.

3.323 As the number of paid individuals that need training is likely to depend on the volume of content that needs to be assessed, the costs of this measure are likely to increase with the benefits to users.

3.324 These costs are also mitigated by the fact that this measure does not specify exactly how providers should provide training to individuals working in search moderation, giving services some flexibility in what they do. Providers can decide the most appropriate and proportionate approach to training such staff for their own contexts. This flexibility allows an approach that is cost-effective and proportionate for each service.

[462] This is based on our assumptions on wage rates as set out in Annex 5. We also assume that the wage cost of the people being trained represents only half of the total costs of the training. Other costs included preparing the training materials, running the training and any related travel to the training. This is consistent with the Department for Education saying that the wage cost of the people being trained accounted for about half of all training expenditure in 2019, although this varies by size of firm and sector. Source: Employer skills survey 2019: Training and Workforce Development - research report, pp. 38 and 40. [accessed: 25 November 2024]. Note that the cost estimate for this measure in the November 2023 Consultation excluded the 22% uplift that we have assumed elsewhere for non-wage labour costs, but we have included this in this updated estimate due to a better understanding of the data. We have also updated these figures since the November 2023 Consultation in line with the latest wage data released by the Office for National Statistics (ONS).

### Rights impact

3.325 This measure should be seen as part of a package of measures relating to search moderation for illegal content, including measures ICS C1 and ICS C2, for which we have assessed the rights impacts above. We do not consider that the measure will have any additional negative impact on these rights.

3.326 Appropriately training individuals involved in search moderation is likely to have significant positive impacts on the rights of users, website or database operators, and service providers, because mistakes are less likely, and moderators will understand their privacy and data protection obligations, where relevant. It will in particular minimise the risk of content being incorrectly reported to reporting authorities in line with their duty under section 66 of the Act (when brought into force) or any other reporting arrangements they have in place, as outlined in paragraph 3.108 in relation to measure ICS C1. To the extent that this measure helps to reduce harm on the service and make users feel safer, this could also positively impact on their human rights.

### Who this measure applies to

3.327 As discussed in our November 2023 Consultation, measure ICS C6 applies where service providers have internal policies in compliance with measure ICS C2. Therefore, it applies only to providers of large general search services or search services identifying risks of multiple harms as only these services are in scope of ICS C2.

3.328 As discussed in 'Our approach to developing Codes measures', we may consult in 2025 on extending this measure to some or all single-risk services.

## Conclusion

3.329 The ability of individuals working in search moderation to deal with illegal content effectively will be materially improved if service providers train them adequately. Where teams are not adequately trained, there is a material risk that they will fail to make appropriate judgments about whether search content is illegal or not. This could result in failure to take action against illegal content, which would increase harm to users. It could also result in over-moderation of legal content, which would be detrimental to freedom of expression.

3.330 As we have shown in the 'Costs and risks' section, the costs of our decision to include provisions around the training of search moderation teams are potentially significant. Nonetheless, given how important effective moderation is to achieving good outcomes for UK users, and given that moderation teams are only likely to be effective where they are adequately trained, we consider this measure to be proportionate.

3.331 Multiple aspects of the decision we are taking reduce the chances of it imposing a disproportionate burden. Firstly, at this time we are focusing the measure on providers of large search services and providers of multi-risk search services. The benefits of such services applying the measure are likely to be particularly high given the greater risks associated with them – and in the case of large services, the number of people that use them in the UK. Secondly, the measure is designed to be flexible and allows service providers to take a tailored approach to training. For example, rather than being prescriptive about what harms providers train their moderation teams in, we have specified that decisions around what training to provider should be informed by providers' risk assessments and by analysis of where there are gaps in expertise amongst their people. This will allow providers to focus their resources on training individuals in issues related to the harms which users of their services are most at risk of being exposed to.

3.332 The full text of the measure can be found in the Illegal Content Codes of Practice for search services and is referred to as ICS C6. This measure is part of our Codes of Practice on terrorism, CSEA and other duties.

# 4. Automated content moderation

## What is this chapter about?

Services use automated tools, often in tandem with human oversight, to make content moderation processes more effective at identifying and removing illegal content or content in breach of their terms of service. As these tools allow services to identify large volumes of harmful content more quickly, they are critical to many services' attempts to reduce harm. This chapter sets out the automated content moderation measures we are recommending, why we are recommending them, and to which user-user (U2U) services they should apply.

## What decisions have we made?

We are recommending the following measures:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| **ICU C9** | Providers should ensure that **hash-matching technology is used** to detect and remove child sexual abuse material (CSAM). This involves analysing images and videos communicated publicly on the service and comparing a digital fingerprint of that content to digital fingerprints of previously identified CSAM. | • Providers of large U2U services which are at medium or high risk of image-based CSAM.<br>• Providers of U2U services which are at high risk of image-based CSAM and have more than 700,000 monthly active UK users.<br>• Providers of U2U services which are at high risk of image-based CSAM and are file-storage and file-sharing services. |
| **ICU C10** | 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**. | • Providers of large U2U services which are at medium or high risk of CSAM URLs.<br>• Providers of U2U services which have more than 700,000 monthly active UK users and are at high risk of CSAM URLs. |

We have also decided **not to recommend** a measure for providers to use standard keyword detection to identify content that is likely to amount to a priority offence concerning articles for use in frauds.

## Why have we made these decisions?

The circulation of CSAM online is increasing rapidly. Child sexual abuse and the circulation of CSAM online causes significant harm, and the ongoing circulation of this imagery can re-traumatise victims and survivors of abuse. Hash matching and URL detection can be useful and effective tools for combatting the circulation of CSAM. While the decisions we are taking today will impose significant costs on some services, we consider these costs are justified given the very serious nature of the harm they address. To ensure that the costs are proportionate, we propose targeting these measures at services where there is a medium or high risk of image-based CSAM or CSAM URLs.

In principle, we consider that, even where they are very small, it would be justified to recommend that services which are high-risk deploy these technologies. However, we have decided to set user-number thresholds below which services will not be in scope of the measure. This is because to implement hash matching and URL detection services will need access to third party databases with records of known CSAM images and lists of URLs associated with CSAM. There are only a limited number of providers of these databases, and they only have capacity to serve a finite number of clients. Setting the user-number thresholds we have should ensure that the database providers have capacity to serve all services in scope of the measure. Should the capacity of database providers expand over time, we will look to review whether the proposed threshold remains appropriate. The evidence we have assessed shows that file sharing services play a particularly significant role in the sharing of CSAM. Therefore, we have decided that all high risk file sharing services should be in scope of our hash matching measure regardless of size.

In the November 2023 Consultation, we proposed recommending the use of standard keyword detection to identify content likely to amount to a priority offence concerning articles for use in frauds to providers of large user-to-user services which are at medium or high risk of fraud. We acknowledged that there is a range of more sophisticated automated tools that service providers may use to detect harmful content (e.g. natural language processing or machine learning). However, we did not have sufficient evidence on the costs and efficacy of these alternative tools to justify recommending their use. Having assessed relevant stakeholder feedback to the consultation, we have decided not to proceed with the measure at this stage. We are instead focusing our efforts on exploring a broader and more flexible measure regarding the use of automated content moderation technologies (including AI), on which we intend to consult in Spring 2025.

# Introduction

The Online Safety Act 2023 (‘the Act’) imposes duties on providers of regulated user-to-user (‘U2U’) services to operate the service using proportionate systems and processes designed to minimise the length of time for which any priority illegal content is present, and to take or use proportionate measures relating to the design or operation of the service to prevent individuals from encountering priority illegal content by means of the service. [463] However, given the volume of user-generated content on many services, human moderation alone is not capable of identifying and removing priority illegal content at sufficient speed and scale. Automated moderation systems and processes are a solution to this issue.

[463] Section 9(2)(a) and (3)(a) of the Act.

4.1 Automated moderation technology can support the identification and removal of priority illegal content, either when it is uploaded or once it is on a service. This includes tools that can compare each piece of content against a database or list of known illegal content using methods such as hash-matching, URL detection, or text detection. Any content that matches existing content in such a list or database can then be flagged for further review or automatically removed. As technology advances, more sophisticated automated technology is becoming more widely available. For example, artificial intelligence can be used to detect first-generation content, which refers to a child sexual abuse image shared for the first time (and which therefore cannot be detected using hash matching).

4.2 At this stage, our approach focuses on the use of automated content moderation (ACM) systems that operate by detecting matches for known child sexual abuse material (CSAM). In particular, this chapter explains our decisions to include measures recommending the use of hash-matching to detect and remove CSAM and the use of technology to detect content matching links or URLs at which CSAM is present in the CSEA Code of Practice. Together, these measures can play an important role in tackling the prevalence and dissemination of CSAM on U2U services.

4.3 We intend to consult on additional measures in spring 2025. This will include work we announced earlier this year, to consult on how automated tools can be used to proactively detect illegal content and the content most harmful to children, going beyond the automated detection measures we are recommending in this chapter.[464]

## Proactive technology measures

4.4 These measures are what the Act describes as "proactive technology measures". The Act places constraints on our power to include such measures in the Codes, including by specifying that such a measure may not recommend the use of 'proactive technology' to analyse user-generated content that is communicated privately (or metadata related to such content). Consistent with that constraint, each of the measures described in this chapter applies only to content communicated publicly by means of the service. We have published guidance on content communicated 'publicly' and 'privately' to assist providers in determining which content on their service is communicated publicly or privately.[465] 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,[466] and there may be good reasons for them to choose to do so in some cases.

4.5 The Act also requires us to have regard to the degree of accuracy, effectiveness and lack of bias achieved by a proactive technology. These matters, in turn, affect the potential impact of proactive technology measures on users, including users' right to freedom of expression and users' privacy. Our assessment of these factors is explained in the relevant sections setting out our reasoning below.

4.6 Our approach aims to set out our recommended measures in sufficient detail to ensure that they are effective and that service providers are readily able to adopt them using any appropriate specific technology or input, providing an appropriate level of flexibility.

[464] Ofcom, "Implementing the Online Safety Act: progress update", October 2024.

[465] 'Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act'.

[466] See also Volume 3, chapter 4: 'Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act'.

## Structure of this chapter

4.7 In our November 2023 Illegal Harms Consultation ('November 2023 Consultation'), we proposed three automated content moderation measures:

a) a recommendation that certain types of services should use an automated technique known as **hash-matching** to analyse images and videos communicated publicly to assess whether they are CSAM and take appropriate measures to swiftly take down CSAM detected. We set out this measure, the impact assessment, and final recommendation from paragraph 4.12;

b) a recommendation that certain types of services should use an automated technique known as **URL detection** to analyse text content communicated publicly to assess whether it consists of or includes CSAM URLs and take appropriate measures to swiftly take down those URLs detected. We set out this measure, the impact assessment, and final recommendation from paragraph 4.206; and

c) a recommendation that services should put in place **keyword detection** technology to identify content that is likely to amount to a priority offence concerning articles for use in frauds. We set out our decision not to recommend this measure from paragraph 4.330.

4.8 In this chapter, we set out and explain our decisions to include (or not include) measures in the Illegal Content Codes of Practice for U2U services relating to the automated moderation of user-generated content. This includes detailing what we proposed in our November 2023 Consultation, the stakeholder feedback on the proposed measures, our decisions and our reasoning.

4.9 This chapter will therefore begin with a discussion of our proposed measures to recommend services use automated techniques known as hash matching and URL detection to analyse relevant content to assess whether it is CSAM and take appropriate steps to remove this content.

4.10 We then review the feedback on the measure on the use of standard keyword detection technology relating to articles for use in frauds, which we have decided not to recommend at this time.

# Measure on using hash-matching to detect and remove CSAM

4.11 As mentioned above (paragraph 4.8 (a)), in our November 2023 Consultation, we proposed a measure recommending that providers of certain U2U services use hash-matching technology effectively to detect known CSAM in the form of images or videos that are (or would be) communicated publicly by means of the service, and swiftly take this content down, for the purpose of complying with their illegal content safety duties under section 10(2) and (3) of the Act.[467] [468]

4.12 The measure proposed the use of a technique called hash-matching, which is a process that detects uploaded (or reuploaded) content that has previously been identified as illegal or otherwise violative. 'Hashing' is an umbrella term for techniques used to create fingerprints

[467] Refer to paragraph 4.12 and 4.13 of this chapter for definition of hash-matching.

[468] See paragraph 4.5.

of content online or files on a computer system. An algorithm known as a hash function is used to compute a fingerprint, known as a hash, from a file.

4.13 Hashes are then stored in a database. These can be used by providers, who generate hashes of the content on their service and compare those against the hashes in the database to test whether any uploaded content is a 'match' for those images. Hash matching can be used to prevent the sharing of illegal or harmful content.

4.14 There are several types of hash matching. Our proposed measure recommended the use of "perceptual" hash-matching over "cryptographic" hash matching, to allow for more harmful content to be identified and potentially moderated. Perceptual hash matching aims to identify images that are similar to images of known CSAM, as opposed to cryptographic hash matching which identifies identical images. In practice, perceptual hash matching is therefore more likely to detect a larger amount of CSAM compared to other forms of hash matching.

4.15 We proposed the measure would apply to the following types of service providers: (1) large services which are at medium or high risk of image-based CSAM in their risk assessment; (2) other services which are at high risk of image-based CSAM in their risk assessment and have more than 700,000 monthly United Kingdom users; and (3) services which are at high risk of image-based CSAM and which are file-storage and file-sharing services that have more than 70,000 monthly UK users.

4.16 We proposed this measure in part because the online circulation of CSAM causes serious and potentially lifelong harm, including re-traumatising victims and survivors of sexual abuse[469] and because our Register of Risks ('Register'), indicates that certain kinds of services and functionalities increase the risk of CSAM offences.[470]

## Summary of stakeholder responses

4.17 We received responses from stakeholders regarding our proposed approach to this measure, which we outline in the following section.

4.18 A range of stakeholders, including providers of regulated services, governments and law enforcement, academics and civil society organisations, expressed broad support for our proposed measure.[471] They agreed that automated moderation systems and processes are

[469] Lee, H. E., Ermakova, T., Ververis, V., and Fabian, B., 2020. Detecting child sexual abuse material: A comprehensive survey. [accessed 22 October 2024]

[470] See Register of Risks chapter Child Sexual Exploitation and Abuse (CSEA).

[471] Are, C. response to November 2023 Illegal Harms Consultation, p.8; Barnardo's response to November 2023 Illegal Harms Consultation, p.16; British and Irish Law, Education, and Technology Association (BILETA) response to November 2023 Illegal Harms Consultation, p.11; Canadian Centre for Child Protection (C3P) response to November 2023 Illegal Harms Consultation, pp.17, 19; Centro de Estudios en Libertad de Expresion y Acceso a la Informacion (CELE) response to November 2023 Illegal Harms Consultation, p.9; Children's Commissioner response to November 2023 Illegal Harms Consultation, p.21; Information Commissioner's Office (ICO) response to November 2023 Illegal Harms Consultation, p.2; International Justice Mission (IJM) response to November 2023 Illegal Harms Consultation, p.15; Internet Matters response to November 2023 Illegal Harms Consultation, p.16; Internet Watch Foundation (IWF) response to November 2023 Illegal Harms Consultation, pp.8, 31; Match Group response to November 2023 Illegal Harms Consultation, p.10; Meta response to November 2023 Illegal Harms Consultation, annex, p.6; Microsoft response to November 2023 Illegal Harms Consultation, pp.12-13; [✂]; National Society for the Prevention of

effective at identifying and removing CSAM from providers' services (with the appropriate design features and safeguards implemented). Furthermore, they agreed that the use of automated moderation tools is necessary due to the volume of CSAM and the seriousness of the harm it causes. These stakeholders also agreed that human moderation is not capable of effectively identifying and removing CSAM on services due to the volume of this content.

4.19 However, several stakeholders commented on – and, in some cases, expressed concerns about – elements of the proposed measure, including:

- the effectiveness of the measure;
- the costs and risks associated with implementing the measure;
- the impacts on users' rights;
- who the measure applies to; and
- our approach to addressing CSAM and the scope of this measure.

4.20 A number of these stakeholders proposed changes to help resolve these concerns. We outline the relevant points and suggestions raised by stakeholders in the following sub-sections.

## Feedback on the measure's effectiveness

4.21 Some stakeholders expressed concerns around the effectiveness of the proposed measure.[472]

4.22 In particular, stakeholders suggested that our approach needed to be future-proof and neutral as to which technologies are recommended, to ensure its long-term applicability. Stop Scams UK advocated for a "dynamic and flexible" approach so that "platforms remain equipped to combat evolving forms of online harm effectively".[473] Meta suggested that technology can evolve very quickly, and said that it "recommend[s] avoiding the imposition of a specific technology". Instead, it suggested we support "sharing of best practices" and allow providers to choose from a "variety of solutions" to make use of "the most up to date

Cruelty to Children (NSPCC) response to November 2023 Illegal Harms Consultation, p.23; Nexus response to November 2023 Illegal Harms Consultation, p.13; OnlyFans response to November 2023 Illegal Harms Consultation, p.5; Scottish Government response to November 2023 Illegal Harms Consultation, p.7; Segregated Payments Ltd response to November 2023 Illegal Harms Consultation, p.8; SPRITE+ (York St John University ) response to November 2023 Illegal Harms Consultation, p.10; Stop Scams UK response to November 2023 Illegal Harms Consultation, pp.10-11; South West Grid for Learning (SWGfL) response to November 2023 Illegal Harms Consultation, p.14; The Independent Inquiry into Child Sexual Abuse (IICSA) Changemakers response to November 2023 Illegal Harms Consultation, p.4; Welsh Government response to November 2023 Illegal Harms Consultation, p.3; WeProtect Global Alliance response to November 2023 Illegal Harms Consultation, p.13; Wikimedia Foundation response to November 2023 Illegal Harms Consultation, p.23; X response to November 2023 Illegal Harms Consultation, p.4.

[472] Barnardo's response to November 2023 Consultation, pp.16-17; Children's Commissioner response to November 2023 Consultation, p.19; Name withheld 5 response to November 2023 Illegal Harms Consultation, pp.10-11; Google response to November 2023 Illegal Harms Consultation, pp.41-42; Glitch response to November 2023 Illegal Harms Consultation, p.8; INVIVIA response to November 2023 Illegal Harms Consultation, pp.14-15; IP.rec response to November 2023 Illegal Harms Consultation, pp.3-4; Meta response to November 2023 Consultation, annex, p.9; Microsoft response to November 2023 Consultation, pp.11-13; Proton response to November 2023 Illegal Harms Consultation, p.6; Reddit response to November 2023 Illegal Harms Consultation, p.10; Stop Scams UK response to November 2023 Consultation, p.11.

[473] Stop Scams UK response to November 2023 Consultation, p.11-13.

technologies”.[474] Google suggested that the proposed measures should be amended to allow service providers to use alternative forms of proactive content moderation which are at least “equally effective” in combatting CSAM. It indicated this flexibility would ensure the measures are future-proofed and prevent there being a “perverse incentive for companies not to adopt” more accurate technologies if they have not yet been assured by Ofcom or reflected in Codes.[475] WeProtect Global Alliance also supported a future-proof and technology-neutral approach.[476] The End Violence Against Women Coalition (EVAW) indicated our approach should “reflect a systems-based approach” and was currently disproportionately focused on content takedown.[477] We address these comments in the ‘Benefits and effectiveness’ section.

4.23 A stakeholder argued that service providers needed more guidance on the types of databases that are acceptable to use for this measure. [✂].[478] We address this concern in the ‘How this measure works’ section.

4.24 Similarly, several stakeholders argued that the accuracy of content (i.e. hashes) included in third-party databases is necessary to ensure its effectiveness in identifying and removing CSAM. Microsoft indicated that hash-matching relies on an accurate database of hashed content. However, it said that hash databases provided by third parties often include content that is not CSAM “but may be related to other images that are CSAM”. It stated that reviewing images “requires extensive investment in technology and tooling, personnel, ongoing training, and wellness, among other things.”[479] Reddit also stated that hash-matching is “only as good as the database it relies on” and requires quality control to ensure appropriate content is included.[480] [✂].[481] IP.rec indicated the importance of the quality of content included in the hash database, citing that inaccurate content inclusion could risk the effectiveness of the measure.[482] We address this concern in the ‘How this measure works’ and the ‘Benefits and effectiveness’ section.

4.25 A number of service providers also indicated that human review contributes to ensuring the effectiveness of hash-matching processes. Microsoft stated that it conducts human reviews “when a match is found on a Microsoft service” including if the hash was received from “trusted third parties such as the Internet Watch Foundation”.[483] [✂].[484] Similarly, [✂] identified that “human review is a critical component of [its] content moderation programmes” indicating it is vital to ensuring the “accuracy of automated tools, protecting the privacy rights of users, and vetting external or user reports of potentially violative content”.[485] [✂].[486] On the other hand, the Canadian Centre for Child Protection (‘C3P’) indicated that using quality hash/URL datasets and supplementing with human moderation is particularly important when the content being classified involves CSAM, as human

[474] Meta response to November 2023 Consultation, annex, p.10.

[475] Google response to November 2023 Consultation, pp.41-42.

[476] WeProtect Global Alliance response to November 2023 Consultation, p.14.

[477] End Violence Against Women Coalition (EVAW) response to November 2023 Illegal Harms Consultation, p.3.

[478] [✂].

[479] Microsoft response to November 2023 Consultation, p.13.

[480] Reddit response to November 2023 Consultation, p.11.

[481] [✂].

[482] IP.rec response to November 2023 Consultation, p.5.

[483] Microsoft response to November 2023 Consultation, p.13.

[484] [✂].

[485] Name withheld 5 response to November 2023 Consultation, pp.10, 15.

[486] [✂].

moderators may otherwise overlook it.[487] We address these concerns in the 'How this measure works',  'Benefits and effectiveness' and 'Risks' sections.

4.26 Lastly, several stakeholders expressed concerns about the accessibility of hash databases for service providers. C3P indicated that it and some other organisations make its hash databases easily available and at a low cost for service providers.[488] The Internet Watch Foundation ('IWF') highlighted it provides Image Hash and URL blocking lists that are high quality, reviewed daily, and recognised as a trusted data source.[489] Protection Group International and an individual argued that hash databases should be made free and easily accessible to all service providers within scope of the proposed measure.[490] Microsoft indicated that it has imposed strict criteria on PhotoDNA licensees to mitigate potential risks of misuse by perpetrators which may limit "the overall availability of PhotoDNA".[491] However, the IWF expressed its ability to support industry, including small and medium sized businesses, with compliance with the Codes by facilitating access to hash databases. It also requested formal recognition of its role as a dataset provider. It also pointed out that "it is essential that an element of due diligence remains in place for service providers like us"[492], and indicated that access to its hash list is "currently tightly controlled through strict contractual arrangements which set out how these services may be deployed in accordance with current laws including GDPR. We also conduct strict due diligence checks on companies and individuals wishing to join the IWF as members and take services".[493] We address these comments in the 'Benefits and effectiveness' section and 'Risks' section.

## Feedback on the measure's risks and costs

4.27 Stakeholders raised several concerns regarding the costs and risks associated with the proposed measure, including about:

- the risk of false positives;
- the security vulnerabilities of hash databases;
- the risk of biases in the collection and assessment of suspected hashes; and
- our cost estimates.

### Risk of false positives

4.28 In the November 2023 Consultation, we described the risk that the use of hash-matching technology results in some content being wrongly detected as a 'false positive' but set out our provisional assessment that perceptual hash-matching systems could be configured to be suitably accurate. Stakeholders generally agreed with this assessment, reporting that hash-matching returned no or few false positives. For example, Match Group said it "[does] not believe [it has] seen any false positives".[494] C3P said that 'Project Arachnid', its content-crawling and hashing software, has issued over 39 million removal requests based on hash

[487] C3P response to November 2023 Consultation, pp.27-28.
[488] C3P response to November 2023 Consultation, p.19.
[489] IWF response to November 2023 Consultation, pp.15-16.
[490] Protection Group International response to November 2023 Illegal Harms Consultation, p.8; Name withheld 2 response to November 2023 Illegal Harms Consultation, p.9.
[491] Microsoft response to November 2023 Consultation, p.14.
[492] IWF response to November 2023 Consultation, pp.3, 4, 15, 16.
[493] IWF response to November 2023 Consultation, p.16.
[494] Match Group response to November 2023 Consultation, p.11.

matching and that "false positives have been rare".[495] Online Safety Tech Industry Association (OSTIA) suggested that further consideration of the "nature of false positives" and the actions taken following detection should be considered along with false positive rates to allow a more comprehensive assessment of the rights impacts.[496]

4.29 However, some stakeholders indicated that hash matching can lead to many false positives (or false negatives).[497] More specifically, [✂] indicated that hash-matching technology relies on sensitivity thresholds to detect "perceptual" matches which aim to find images close to the original image but "which in practice can, and do with sufficient regularity, result in false positives".[498] It also indicated the risk of false positives requires human review as it ensures the "accuracy of automated tools, protecting the privacy rights of users, and vetting external or user reports of potentially violative content".[499] Another stakeholder indicated that [✂].[500] INVIVIA suggested that hash-matching is generally accurate, however, there are instances of false positives and false negatives. It suggested that advanced cryptographic protocols and technologies are being developed to reduce these issues and enhance public verification processes.[501] [✂].[502] We address these concerns and comments regarding the risk of false positives in the 'Risks' section.

4.30 The Information Commissioner's Office ('ICO') expressed concern that the measure may "allow for the content moderation technology to be configured in such a way that recognises that false positives will be reported to the NCA".[503] It indicated that it is concerned that a margin for error could be routinely "factored into" a service's systems and processes and recommended that service providers should be explicitly required to take into account the importance of minimising false positives being reported to the NCA.[504] This point is addressed in the 'how this measure works' section.

### Security vulnerabilities of hash databases

4.31 A number of stakeholders expressed concerns about security risks regarding the deployment of hash-matching technology. Proton stated that "maintaining a functioning hash matching technology on a service is difficult and does not come without strong privacy and security risks".[505] Microsoft argued that "one reason commercial viability of hash-matching technologies is limited is because of the potential for misuse by adversarial actors. These adversarial actors have attempted to leverage our hash-matching technology to reverse engineer the content of hash datasets in an effort to circumvent detection of known CSAM and other hashed content areas. For this reason, Microsoft has imposed strict criteria on potential licensees, limiting the overall availability of PhotoDNA so as to remove

[495] C3P response to November 2023 Consultation, p.19.

[496] Online Safety Tech Industry Association (OSTIA) response to November 2023 Illegal Harms Consultation, p.13.

[497] BILETA response to November 2023 Consultation, p.12; Global Partners Digital response to November 2023 Illegal Harms Consultation, p.14; Integrity Institute response to November 2023 Illegal Harms Consultation, p.12; Proton response to November 2023 Consultation, p.6.

[498] Name withheld 5 response to November 2023 Consultation, p.10.

[499] Name withheld 5 response to November 2023 Consultation, p.10.

[500] [✂].

[501] INVIVIA response to November 2023 Consultation, p.11.

[502] [✂].

[503] ICO response to November 2023 Consultation, p.14.

[504] ICO response to November 2023 Consultation, p.14.

[505] Proton response to November 2023 Consultation, p.6.

these potential risks".[506] INVIVIA noted that "services looking to access these databases must comply with specific security and operational standards to ensure the sensitive data is handled appropriately. This includes data protection measures, secure access protocols, and possibly undergoing security audits".[507] Global Partners Digital provided a specific example of potential attacks impacting hash-matching technology, highlighting the potential vulnerabilities of hash databases with public algorithms.[508]We address these comments in the 'Risks' section.

### Risk of biases in the collection and assessment of suspected hashes

4.32 We also received input from OSTIA, Glitch and IP.rec noting the risk of biases in hash databases, which may disproportionately (mis)represent certain groups. OSTIA suggested we explicitly recommend that third-party database providers should avoid "systematic bias within their control", arguing that databases should "determine addition of content solely based on whether or not it is CSAM and ensure minimisation of bias in processes making that determination".[509] Glitch also shared concerns that automated moderation technologies "may perpetuate biases present in training data or design, leading to disproportionate moderation outcomes for women and girls".[510] IP.rec indicated that "the integration of perceptual hash matching technology with machine learning and artificial intelligence tools exponentially multiplies the consequences of errors in the databases, as it increases the likelihood of the system reproducing racist and xenophobic biases, resulting in a high rate of false positives and false negatives".[511] We address these comments in the 'Risks' section.

### Cost estimates

4.33 Some stakeholders outlined concerns regarding the estimated costs of implementing the measure. A few stakeholders commented that the costs associated with reviewing, moderating, and reporting CSAM could be particularly burdensome for providers of small and medium sized services.[512] On the other hand, C3P argued hash-matching technology would not be costly or difficult to deploy for smaller service providers. It suggested that the costs of hash-matching for smaller service providers were likely to be negligible. It also noted that it provides its hash lists to smaller service providers at no cost.[513] Protection Group International indicated that some figures in our cost estimates included in the November 2023 Consultation appeared incorrect, as they did not show fees paid to organisations to supply hashes. It stated that "the figures for actionable content from NCMEC (National Center for Missing & Exploited Children) referrals isn't clear to make a full judgement of costs about safeguarding".[514] It also suggested the cost figures needed to be updated. We consider these comments in the 'Costs' section and address them in further detail in Annex 5.

[506] Microsoft response to November 2023 Consultation, pp.13-14.

[507] INVIVIA response to November 2023 Consultation, p.15.

[508] Global Partners Digital response to November 2023 Consultation, p.14.

[509] OSTIA response to November 2023 Consultation, p.14.

[510] Glitch response to November 2023 Consultation, p.8.

[511] IP.rec response to November 2023 Consultation, p.5.

[512] INVIVIA response to November 2023 Consultation, pp.14-15; Integrity Institute response to November 2023 Consultation, p.12; Scottish Government response to November 2023 Consultation, p.7.

[513] C3P response to November 2023 Consultation, p.19.

[514] Protection Group International response to November 2023 Consultation, p.5.

4.34 Stakeholders also argued that the cost assessment did not consider other important factors. Yoti and IWF argued that the measure's costs for the broader ecosystem involved in child sexual abuse, including law enforcement and civil society, should be considered.[515] The IWF highlighted that there are costs associated with scaling automated moderation services, but there is "uncertainty about the number of organisations that may be creating demand". It also suggested that providers of small and medium services likely to fall within scope of the measure would not generate significant revenue, but may incur several costs [for the IWF] associated to due diligence, contractual work, monitoring, and compliance.[516] Another stakeholder, INVIVIA, suggested that any specific hash database "would need to be further evaluated for scaling capacity and cost impacts at scale".[517] We consider these concerns in the 'Costs' and 'Who this measure applies to' sections.

## Feedback on the impact to users' rights

4.35 Several stakeholders provided input on the potential impacts of this measure on users' rights to privacy and freedom of expression. The Oxford Disinformation and Extremism Lab expressed concerns that the proposed measures will impact specific users' rights to freedom of expression, notably civil society organisations, academic researchers, and human rights advocates. They also indicated the importance of working with academics and public-private partnerships to pursue implementing hash-matching.[518]

4.36 C3P emphasised that "allowing CSAM to be uploaded and shared is a massive violation of the privacy of the person depicted".[519]

4.37 The ICO also noted that the accuracy principle in data protection law "requires that [service providers] take all reasonable steps to ensure that the personal data they process is not incorrect or misleading as to any matter of fact". It observed that "the level of accuracy that is appropriate for reports to the National Crime Agency ("NCA") (which carries a particular risk of serious damage to the rights, freedoms and interests of a person who is incorrectly reported) and other significant but potentially less harmful actions such as content takedown" differed. It suggested that the measure should set out that service providers "take into account the importance of minimising false positives being reported to the NCA" when configuring hash-matching technology and when deciding what proportion of detected content human moderators should review.[520]

4.38 We consider these concerns and comments in the 'Rights impacts' section.

## Feedback on who this measure applies to

4.39 Several stakeholders expressed support for our proposed approach to the types of service providers within scope of the measure. The Marie Collins Foundation agreed that the overall approach of applying more burdensome measures to large and high-risk services was appropriate, and supported our proposal that smaller services which were high-risk for

[515] IWF response to November 2023 Consultation, p.16; Yoti response to November 2023 Illegal Harms Consultation, p.11.

[516] IWF response to November 2023 Consultation, p.16.

[517] INVIVIA response to November 2023 Consultation, p.17.

[518] Oxford Disinformation and Extremism Lab response to November 2023 Illegal Harms Consultation, pp.13-14.

[519] C3P response to November 2023 Consultation, p.18.

[520] ICO response to November 2023 Consultation, pp.13-14.

grooming and CSAM should also be in scope of more onerous measures.[521] The IJM highlighted the availability of free or low-cost hash-matching technology available for services looking to comply with the measure, and argued that it is imperative that even smaller or seemingly less risky companies adhere to every measure outlined in the Codes.[522] WeProtect Global Alliance and C3P were particularly supportive of file-storage and file-sharing services being within the scope of the measure.[523]

4.40 We also received input from OSTIA arguing that our proposed approach made "an implicit assumption that a new risk assessment on the service deemed low risk would identify the presence of CSAM and thus increased risk of CSAM in future". It said that it believed this assumption "is likely to be incorrect in most practical cases".[524] We understand this to express a concern that the provider of a low risk service which does not implement proactive content moderation measures such as hash matching would not become aware of the presence of illegal content on the service and reassess its risk level.

4.41 We address this feedback in the 'Who this measure applies to' section.

### Scope of applicable services

4.42 Comparatively, some stakeholders suggested the scope of the measure should be broadened to include more services or different types of services.[525] The Cyber Helpline indicated that "content moderation codes relating to CSAM should apply to all services in scope".[526] The Marie Collins Foundation stated that "[user number thresholds] should be constantly reviewed" and that "all services that are high or medium risk for image-based CSAM regardless of size need to implement ACM".[527] The IWF proposed that the measure "should apply to all services at medium to high risk of one type of harm", implying that there should not be any user reach threshold for inclusion.[528] However, it suggested that micro, small and medium sized businesses could be given "longer to prepare, maybe a period of 12 to 18 months" for the implementation of this measure.[529] The Children's Commissioner recommended that "child safety measures should be applied to all user-to-user services that children may use in order to avoid loopholes that are exploited by unregulated services".[530] We address this feedback in the 'Who this measure applies to' section.

[521] Marie Collins Foundation response to November 2023 Illegal Harms Consultation, p.8.

[522] IJM response to November 2023 Consultation, pp.10, 11, 13.

[523] WeProtect Global Alliance response to November 2023 Consultation, p.13; C3P response to November 2023 Consultation, p.5.

[524] OSTIA response to November 2023 Consultation, pp.14-15.

[525] Barnardo's response to November 2023 Consultation, pp.10, 13, 16-17; C3P response to November 2023 Consultation, pp.4, 12; Children's Commissioner response to November 2023 Consultation, p.19; Cyacomb response to November 2023 Illegal Harms Consultation, p.11; IWF response to November 2023 Consultation, p.31; Marie Collins Foundation response to November 2023 Consultation, p.8; NSPCC response to November 2023 Consultation, p.24; OSTIA response to November 2023 Consultation, p.10; The Cyber Helpline response to November 2023 Illegal Harms Consultation, p.10.

[526] The Cyber Helpline response to November 2023 Consultation, p.10.

[527] Marie Collins Foundation response to November 2023 Consultation, p.11.

[528] IWF response to November 2023 Consultation, p.31.

[529] IWF response to November 2023 Consultation, p.21.

[530] Children's Commissioner response to November 2023 Consultation, p.19.

4.43 Cyacomb and OSTIA asked us to clarify how often the database capacity should be reviewed to gradually include more services.[531] This is a response to our explanation, in the November 2023 Consultation, that we wanted to avoid a scenario in which the database ecosystem was unable to cope with demand from online services in the short term. The stakeholders suggested we consider when a review would be needed, what it should cover, and who would be affected by it, arguing that "clarity could create the incentive needed for investment in change".[532] We address this concern in the 'Who this measure applies to' section.

4.44 On the other hand, some service providers contended that our proposed measure would apply too broadly. Service providers, such as Booking.com and [✂], commented that the risk of CSAM should be assessed by taking into account the nature of the platform.[533] One individual also argued that it was disproportionate to require the use of hash-matching if a service has no history of storing or distributing CSAM, recommending that the measure should be mandatory if there is evidence that CSAM is being stored or distributed on the service, or the risk assessment indicates there is risk of it.[534] This concern is addressed in the 'Costs' section.

### Smaller services

4.45 Specifically, several stakeholders argued in favour of expanding the scope of the measure to include providers of smaller services. C3P indicated that a range of smaller services used to spread CSAM were not captured by this measure. It also stated that "several smaller companies are using C3P hash lists for free, and there is no specialist knowledge required".[535] Similarly, Barnardo's also suggested that undue focus on proportionality also means that many small companies will be exempt from following many of the proposed measures" and "there is potential to let harmful and/ or risky small companies off the hook – such as collector sites for CSAM, where the risk is high and harmful".[536] The National Society for the Prevention of Cruelty to Children (NSPCC) suggested the measure should be applied to "smaller services (700,000 users) with a medium risk of CSAM (rather than just a high risk)".[537] We address these concerns in the 'Who this measure applies to' section.

4.46 On the other hand, several stakeholders suggested the costs associated with hash-matching would make the measure difficult for providers of smaller services to implement.[538] Microsoft indicated that "though human review is not required, reliance on third-party hashes (even from a trusted source) has costly consequences elsewhere in the ecosystem". It also stated that the current proposals "assume reliable and low-cost technology is commercially available to companies", which it does not believe to be accurate.[539] Similarly,

[531] Cyacomb response to November 2023 Consultation, p.12; OSTIA response to November 2023 Consultation, p.11.

[532] Cyacomb response to November 2023 Consultation, p.12; OSTIA response to November 2023 Consultation, p.11.

[533] [✂]; Booking.com response to November 2023 Illegal Harms Consultation, p.11.

[534] Carr, J. response to November 2023 Illegal Harms Consultation, p.8.

[535] C3P response to November 2023 Consultation, pp.4, 12, 19.

[536] Barnardo's response to November 2023 Consultation, p.10.

[537] NSPCC response to November 2023 Consultation, p.24.

[538] Name withheld 5 response to November 2023 Consultation, p.11; Integrity Institute response to November 2023 Consultation, p.12; INVIVIA response to November 2023 Consultation, pp.14-15; Microsoft response to November 2023 Consultation, p.13; [✂]; Protection Group International response to November 2023 Consultation, p.8.

[539] Microsoft response to November 2023 Consultation, p.13.

INVIVIA, the Integrity Institute and [✂] highlighted that hash-matching technology can be expensive and places a demand on resources, noting that smaller services would need more support in deploying and using the databases. INVIVIA specifically indicated that "beyond setup, there are ongoing costs associated with maintaining the system, updating hash databases, and managing false positives and negatives, which require manual review and can strain limited resources".[540] The Integrity Institute highlighted that providers of small services will struggle "when dealing with reporting and legal aspects".[541] [✂] stated that costs associated with a detection programme (both automated and human) can be "extensive and ongoing" including "technical costs", and "acquisition and quality control of ingested hash sets".[542] [✂].[543] We address this feedback in the 'Who this measure applies to' section.

### File-sharing services

4.47 We received several responses which supported our focus on file-storage and file-sharing service providers. For example, WeProtect Global Alliance expressed its support for applying the measure to file-storage and file-sharing services, noting specifically that "perpetrators of child sexual abuse typically use cloud file sharing to efficiently exchange images and videos with both known and new offender contacts. To ensure that content remains accessible for as long as possible, determined offenders use multiple cloud platforms simultaneously. The true nature of harmful links is hidden behind a smoke screen of references to other (lesser) illegal activity or legitimate file-sharing uses to evade detection".[544] The use of file-storage and file-sharing services by perpetrators to spread CSAM was also highlighted by C3P in its consultation response. Specifically, C3P indicated "…we have countless examples of smaller services being used to share CSAM and are aware that the selection of certain services can be an orchestrated action by the perpetrator community. Looking at the most common file-hosting services used to distribute CSAM based on Project Arachnid data, it is unlikely that any of these services — most of which would not be recognized by average citizens — would meet the 7 million UK user threshold. We strongly recommend considering either a lower user threshold or a threshold based on total bandwidth".[545] C3P also drew attention to a specific example of a file-storage and file-sharing service which was below the 70,000 user threshold we had proposed but (it said) had been used to store and disseminate CSAM at scale.[546]

4.48 We also received feedback from some file-sharing service providers, or providers that had a file-sharing element, which expressed reservations about our proposed approach. These comments related in part to the draft Risk Assessment Guidance which was published as part of the November 2023 Consultation. The draft guidance included a CSAM "risk decision framework" which stated that file-storage and file-sharing services which allowed images or

[540] INVIVIA response to November 2023 Consultation, pp.14-15.

[541] Integrity Institute response to November 2023 Consultation, p.12.

[542] Name withheld 5 response to November 2023 Consultation, p.11.

[543] [✂].

[544] WeProtect Global Alliance response to November 2023 Consultation, 2023, p.13.

[545] C3P response to November 2023 Consultation, p.5. This comment was made in relation to another of our proposals about annual review of risk management activities (which applies to large services). We consider it is also relevant to our hash-matching proposal (which was proposed for a greater range of services).

[546] C3P response to November 2023 Illegal Harms Consultation, p.18.

videos to be uploaded, posted or sent were likely to be high risk for image-based CSAM.[547] Proton (which has a cloud storage service called Proton Drive) set out its view that it was "unfair and disproportionate" to automatically assume file-sharing services were high risk for image-based CSAM "as it does not take into account the potential measures taken by the service to mitigate harm, its number of users nor any past records of misuse".[548] A stakeholder indicated that public file sharing and file storage services should not be considered equivalent for risk assessment purposes, given that public file sharing entails a much higher risk profile than personal file storage and family file-sharing services.[549] [✂].[550]

4.49 We address these concerns and comments in the 'who this measure applies to' section.

### End-to-end encrypted services

4.50 We also received several responses expressing views about applying this measure to end-to-end encrypted services.

- Global Partners Digital indicated that scanning content on encrypted services risks interfering with freedom of expression and the right to privacy, and it might not be technically feasible "without affecting the security of the system as a whole".[551]
- The Electronic Frontier Foundation ('EFF') stated that it "agree[d] with Ofcom's apparent decision to not mandate or encourage scanning of any encrypted communications, since this would constitute a 'backdoor' method of reading private user data". It also indicated "that hash-matching techniques should not be applied to encrypted data (including client-side scanning techniques that scan data before or after the encryption algorithm is applied)".[552]
- Wikimedia Foundation indicated the proposed measure should not discourage or prohibit the use of end-to-end encrypted communications or de-incentivise platforms and other service providers from offering them to safeguard the privacy and safety of their users.[553]
- Global Network Initiative ('GNI') cautioned against the over-reliance on automated detection technologies that could disincentivise the adoption of encrypted technologies and suggested creating an exception from any scanning obligations for encrypted services.[554]

4.51 Conversely, several civil society organisations argued that the measure should apply to end-to-end encrypted services. The Marie Collins Foundation and WeProtect Global Alliance argued that it was important to apply this measure to end-to-end encrypted services because CSAM is very prevalent on these services.[555] The Independent Inquiry into Child Sexual Abuse ('IICSA') Changemakers indicated the proposed measure should apply to end-to-end encrypted services, stating that they should be compelled to enable law

[547] As explained in the proposed Risk Assessment Guidance in Table 7 as part of our November 2023 Consultation.

[548] Proton response to November 2023 Consultation, p.5.

[549] [✂].

[550] [✂].

[551] Global Partners Digital response to November 2023 Consultation, p.14.

[552] The Electronic Frontier Foundation (EFF) response to November 2023 Illegal Harms Consultation, pp.1-2.

[553] Wikimedia Foundation response to November 2023 Consultation, p.23.

[554] Global Network Initiative response to November 2023 Illegal Harms Consultation, p.11.

[555] Marie Collins Foundation response to November 2023 Consultation, p.11; WeProtect Global Alliance response to November 2023 Consultation, p.14.

enforcement agencies to detect CSAM and support enquiries.[556] The IJM argued that there was an urgent need for service providers to adopt advanced technological solutions that proactively identify and mitigate the dissemination of harmful content.[557] Lastly, the Phoenix 11 said that solutions to detect CSAM in end-to-end encrypted environments do exist and should be applied to these services, citing the example of technology developed by Apple to detect known CSAM.[558]

4.52 Among other things, these comments relate to the question of whether it is 'technically feasible' for a service provider to implement the measure, and the distinct question of whether content should be considered as communicated 'publicly' or 'privately'. We address the first question in the 'How the measure works' section (paragraph 4.70). As to the second question, we explain in Volume 3, chapter 4 ('Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act') that we do not agree with stakeholders who argued that content communicated on end-to-end encrypted parts of a service should always be considered to be communicated 'privately' for the purposes of the Act. Service providers should refer to the guidance for assistance in determining whether content should be considered as communicated 'publicly' or 'privately'. [559]

4.53 We address these concerns and comments in the 'who this measure applies to' section.

## Feedback on the scope of this measure and what it applies to

4.54 Some stakeholders responded to the proposed approach for tackling CSAM, in particular that the measure only facilitates detection of known CSAM and is narrowly focussed on one type of illegal content.

4.55 A specific concern raised was that using hash-matching alone to tackle CSAM is not sufficient because the technology cannot detect new CSAM. [✂].[560] The International Justice Mission Center to End Online Sexual Exploitation of Children (IJM) also raised concerns that overemphasising the detection of known CSAM might incentivise the production of new CSAM (including via livestreaming).[561] We address these concerns in the 'Effectiveness' section.

4.56 In a similar vein, several stakeholders suggested alternative systems and processes to identify new CSAM. The Age Verification Providers Association (AVPA) and VerifyMy indicated the measure should set out a requirement to apply automated age assurance to detect newly generated CSAM.[562] The UK Safer Internet Centre (UKSIC) suggested a requirement for service providers to deploy classifiers to detect new CSAM[563] and Internet Matters recommended establishing an evidence base on this type of classifier technology.[564] We address this in the 'How this measure works' section.

[556] IICSA Changemakers response to November 2023 Consultation, p.3.

[557] IJM response to November 2023 Consultation, p.3.

[558] Phoenix 11 response to November 2023 Illegal Harms Consultation, p.2; Apple, 2021. CSAM Detection Technical Summary.

[559] Guidance on content 'communicated 'publicly' and 'privately' under the Online Safety Act.

[560] [✂].

[561] IJM response to November 2023 Consultation, p.4.

[562] Age Verification Providers Association (AVPA) response to November 2023 Illegal Harms Consultation, p.2; VerifyMy response to November 2023 Illegal Harms Consultation, p.8.

[563] UK Safer Internet Centre (UKSIC) response to November 2023 Illegal Harms Consultation, pp.16, 39.

[564] Internet Matters response to November 2023 Consultation, p.16.

4.57 Lastly, several stakeholders provided input on the scope of the measure, and specifically about expanding the kinds of harms the measure applies to. Several stakeholders provided evidence on applying hash-matching to detect and remove terrorism content: [565]

- Tech Against Terrorism indicated the efficacy and utility of hash-matching for detecting and removing terrorism content, although it noted important considerations around contextualising content and implications on users' rights, such as freedom of expression.[566]
- The New Zealand Classification Office indicated the Global Internet Forum to Counter Terrorism's hash-sharing database is an example of effective tools for hash-matching terrorism content.[567]
- Comparatively, the Electronic Frontier Foundation (EFF) expressed concerns about the use of hash-matching technology to detect and remove terrorism content, indicating that there is a lack of transparency and oversight in using this technology.[568]
- Microsoft also encouraged Ofcom to carefully consider requiring service providers to deploy hash-matching (and URL detection) for terrorism content.[569]

4.58 Several other stakeholders suggested broadening the approach to tackle other types of illegal content:

- Glitch indicated the measure does not address potential mitigations for gender-based harms, including its intersectionality with CSAM.[570]
- Refuge queried why the measure did not apply to adult intimate image abuse.[571]
- UK Safer Internet Centre (UKSIC) recommended that the measure require service providers to use hashes from StopNCII.org.[572]
- Four Paws indicated the measure should not extend to the removal of educational content on animal welfare.[573]
- SWFGfL recommended the measure apply to other types of illegal harm.[574]

4.59 We address these comments in the 'Effectiveness' section.

## Our decision

4.60 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation.

4.61 We have made a change to the measure by **removing the user threshold for file-storage and file-sharing services, so that the measure applies to all such services at high risk of**

[565] In the November 2023 consultation, we invited evidence on applying hash matching and URL detection for terrorism content to a range of services.
[566] Tech Against Terrorism response to November 2023 Illegal Harms Consultation, pp.15-16.
[567] New Zealand Classification Office response to November 2023 Illegal Harms Consultation, p.8.
[568] Electronic Frontier Foundation (EFF) response to November 2023 Illegal Harms Consultation, p.14.
[569] Microsoft response to November 2023 Consultation, p.14.
[570] Glitch response to November 2023 Consultation, p.8.
[571] Refuge response to November 2023 Illegal Harms Consultation, p.13.
[572] UKSIC response to November 2023 Consultation, p.3.
[573] Four Paws response to November 2023 Illegal Harms Consultation, p.14.
[574] SWFGfL response to November 2023 Consultation, p.14.

**image-based CSAM**. Having reviewed the evidence of harm from CSAM on file-storage and file-sharing services, and considering the consultation responses, we have decided not to include the user threshold we had proposed (of 70,000 monthly United Kingdom users) for this service type.

4.62 Our measure therefore sets out that providers of certain types of service should use an automated technique known as hash matching to analyse relevant content to assess whether it is CSAM and should take appropriate measures to swiftly take down CSAM detected. This measure applies to:

- large services which are at medium or high risk of image-based CSAM
- services which are at high risk of image-based CSAM and have more than 700,000 monthly active United Kingdom users[575](‘monthly UK users’)
- services which are at high risk of image-based CSAM and are file-storage and file-sharing services

4.63 We have also made a number of minor amendments to the measure in response to the feedback we received. These changes are:

- **Clarifying the scope of the hashes to be used:** We have clarified that hashes should be sourced from at least one organisation with expertise in the identification of CSAM, and that service providers can source hashes from more than one such organisation. They can also use other hashes of CSAM, such as CSAM identified by the service’s content moderation function. In addition, the measure sets out that the set of hashes in use should (taken together) reflect the range of child sexual abuse material that is illegal in the UK.
- **Reducing the risk of bias:** We have made a change to ensure that the arrangements operated by the organisation(s) from which hash databases are sourced for identifying or assessing suspected CSAM do not plainly discriminate on the basis of protected characteristics (such as sex or race). This change is intended to address potential risks relating to bias.
- **Reducing the risk of reporting false positives:** We have added the importance of minimising the reporting of false positives to the NCA or a foreign agency (such as NCMEC) as a matter that service providers should take into account when configuring the hash matching technology (and its balance between precision and recall), and when deciding the proportion of detected content that should be reviewed by human moderators.[576]
- **Securing hash databases:** We made a change by specifying that service providers should ensure that an appropriate policy is in place, and measures are taken in accordance with that policy, to secure any hashes of CSAM from unauthorised access, interference, or exploitation. We also made a change to ensure that the organisation(s) from which hash databases are sourced secure those databases.
- **Safeguarding users’ rights:** We have specified which other measures in our Codes act as safeguards for users’ right to freedom of expression and privacy. These include measures relating to content moderation, complaints procedures, and terms of service.

[575] As calculated in accordance with the methodology set out in the Codes of Practice. See the ‘Our approach to developing Codes measures’ chapter for more information.

[576] Precision is the proportion of positive classifications that were correctly identified as positive. Recall is the proportion of all positives that were classified as positive.

4.64 The measure can be found in our Illegal Content Codes of Practice for U2U services and is referred to as ICU C9. This measure is part of our CSEA Code.

# Our reasoning

## How this measure works

4.65 This measure sets out that providers of certain types of service should use perceptual hash-matching technology to detect and remove child sexual abuse material (CSAM). This involves analysing images and videos communicated publicly on the service and comparing a digital fingerprint of that content to digital fingerprints of previously identified CSAM.

### Use of hash-matching technology

4.66 The measure sets out how service providers should use hash-matching technology effectively to detect CSAM. It recognises that the set of hashes used is critically important to the overall functioning of perceptual hash-matching technology and is an essential element to ensuring the effectiveness of this measure.

4.67 In particular, the measure specifies that the set of hashes used should include at least one hash database sourced from an appropriate organisation. We consider that there are several conditions that any hash database should meet for it to be considered appropriate for use by a service provider. We designed these conditions to help ensure that the adoption of hash-matching technology is effective and accurate in mitigating against the circulation of CSAM on a service and includes appropriate safeguards to mitigate potential rights impacts or misuse of the technology. The conditions also reflect the constraints that the Act places on "proactive technology measures", where we are required to have regard to the degree of accuracy, effectiveness and lack of bias achieved by a proactive technology (see paragraph 4.5).

4.68 We detail these conditions below.

4.69 **Providers should use perceptual hash-matching technology to analyse content in the form of images or videos which is communicated publicly on the service, and should take appropriate measures to swiftly take down (or prevent from being generated, uploaded, or shared) detected content that is CSAM (subject to human moderation, as detailed below).**

a) As explained in paragraph 4.5, the measure only applies in relation to content communicated publicly on the service, consistent with constraints on our powers to recommend the use of 'proactive technology'.

b) Our recommendations in this chapter will also only apply where it is technically feasible for a provider to implement them. 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, our measures do not apply to providers that are technically unable 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.

c) The measure sets out that relevant content already present on the service at the time the technology is implemented should be hash-matched within a reasonable time. New content generated, uploaded to, or shared on the service after the technology is implemented should be hash-matched before (or as soon as practicable after) it can be encountered by UK users of the service.

4.70 **Providers should compare content to an appropriate set of hashes of known CSAM using a suitable perceptual hash function, configured to strike an appropriate balance between precision and recall.**

a) The measure sets out that the hashes used should include hashes of CSAM sourced from one or more persons/organisations with expertise in the identification of CSAM and who meet other requirements designed to ensure the database provided is accurate and effectively maintained. We refer to such persons/organisations as an organisation for ease of reference. These requirements include (1) that the organisation has arrangements in place to identify suspected CSAM and secure that as far as possible) it is correctly identified before hashes are added to the database; (2) that these arrangements do not plainly discriminate on the basis of protected characteristics such as sex or race; (3) that the organisation also has arrangements in place to regularly update its database with newly identified hashes of CSAM, and (4) to review cases where material is suspected to have been incorrectly identified and remove such hashes from the database where appropriate; and (5) that the organisation has arrangements in place to secure the database from unauthorised access, interference or exploitation. We amended the measure to clarify that service providers can source hashes from more than one such organisation.

b) The measure further clarifies that service providers may use other CSAM hashes, such as hashes of material identified by the service's own content moderation function. When using this source of hashes, the provider must secure as far as possible that CSAM is correctly identified and review and remove instances where material has been incorrectly identified as CSAM.

c) The measure also sets out that the service provider should ensure that the set of hashes used, taken together, reflects the range of CSAM that is illegal under criminal law in the UK (including images of children that are indecent and illegal but may not show sexual activity).[577] This reflects that the definition of CSAM in the UK is broader than the criminal law of other jurisdictions (and for instance extends to non-photographic child sexual abuse imagery).[578] We explain this change further in the effectiveness section at paragraphs 4.93 to 4.95.

d) The measure also sets out that the perceptual hashing matching technology should be configured so that its performance strikes an appropriate balance between precision and recall. It describes factors that the provider should take into account when configuring the technology. These factors include the service's risk of harm relating to image-based CSAM[579] and the proportion of detected content that is a "false positive" match to the hash database/s that are used. The provider should keep a written record

[577] This includes in particular "Category C" images (as referred to in the guidelines issued by the Sentencing Council for England and Wales). Category C refers to images that do not depict sexual activity falling within Categories A or B but are indecent and illegal under the relevant law.

[578] Information about what content amounts to CSAM can be found in Ofcom's Illegal Content Judgements Guidance.

[579] This should reflect the findings of the service's latest illegal content risk assessment and any information reasonably available to the service provider about the prevalence of CSAM within the content to which the measure applies. Such information could include takedown notices received from NGOs.

of its approach to configuring the technology and review it at least every six months. We discuss this further in the risks section at paragraphs 4.128 to 4.133.

e) The measure also sets out that service providers should ensure an appropriate policy is in place, and security measures taken in accordance with that policy, to secure any hashes of CSAM held for the purposes of the measure (including any copy of a hash database sourced from an appropriate organisation). This is to protect against unauthorised access, interference, or exploitation. Where the provider holds a copy of a hash database sourced from an appropriate organisation, such security measures will often be a contractual requirement. We slightly strengthened this provision to include the need for a policy to be put in place, to promote good decision-making about which measures to implement. These should include technical and non-technical measures (comprising of a mix of procedural, physical, personnel, and technical controls) to secure against adversarial attacks and exploitation.

4.71 We consider best practices for mitigating security risks may include, but are not limited to:

- storing data securely within the service's systems;
- restricting access to the CSAM hash database to authorised persons only;
- maintaining records of all authorised persons;
- ensuring all authorised persons have an appropriate understanding of how the measure operates;
- requiring multifactor authentication for access to an account capable of making changes to the CSAM hash database;
- requiring that changes to the CSAM hash database (or how the measure is implemented) must be proposed and approved by more than one authorised person;
- retaining records of (1) all changes to the CSAM hash database, (2) changes to how the measure is implemented, and (3) the authorised person(s) who proposed and approved any changes;
- avoiding the use of default or shared passwords and credentials for accounts providing access to the CSAM hash database; and
- ensuring that passwords and credentials are managed, stored, and assigned securely, and are revoked when no longer needed.

### Use of human moderators and record-keeping

4.72 The measure sets out how service providers should ensure they have appropriate policies for reviewing detected content, and have processes for keeping statistical records.

4.73 **Providers should put in place a policy for review of content detected by the hash-matching technology which ensures that an appropriate proportion of detected content is reviewed by human moderators, and act according to this policy**. This is a safeguard to identify false positives and limit adverse impacts on users' rights. The measure enables the service provider to decide what proportion to review, but specifies things that should be taken into account in deciding its policy:

- 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
- the importance of minimising the reporting of false positives to the NCA or a foreign agency. The need to take steps to reduce the amount of false positives reported to the

NCA or a foreign agency was raised by stakeholders during the consultation (refer to paragraph 4.29 and 4.31).

4.74 **Providers should ensure that a written record is made of their policies for review, setting out the proportion of content which is intended to be reviewed and information about how the things set out above were taken into account.** This is designed to promote good decision-making.

4.75 **Providers should keep statistical records about content reviewed (including the number of reviews carried out, the proportion of detected content this represents, and the number of false positives identified).**

- These records can then be used in the periodic reviews of the performance of the hash-matching technology, together with other data from the service's complaints procedure (where users complain that their content has been wrongly identified as illegal content).
- In turn, the provider can adjust its policy for review of detected content, as appropriate.

4.76 Our approach to human moderation is substantially the same as we proposed in the November 2023 Consultation. We have made clearer that service providers should put in place a policy for review. We have also added the importance of minimising the reporting of false positives to the NCA or a foreign agency (such as NCMEC) as a matter that service providers should take into account when deciding on their policy, as suggested by the ICO (see paragraph 4.38).

4.77 A stakeholder commented on the potential for increased viewing of CSAM by human moderators and the risks associated with exposure to this material.[580] We recognise that exposure to CSAM could be detrimental to moderators' wellbeing and would encourage service providers to take appropriate steps to support those working with CSAM.

4.78 Our approach also recognises that service providers may have systems and processes other than human review in place to minimise false positives. For example, a service provider might:

- use cryptographic hash-matching to identify if an item of content detected as a match by perceptual hash-matching technology was an exact match for known CSAM;
- use more than one perceptual hash-matching algorithm (reducing the likelihood that each algorithm results in a false positive for a particular item of content considerably); and
- using machine-learning classifiers to identify items of detected content that are more or less likely to be false positives.

4.79 Where human moderators do not review all detected content, we consider best practices for review of detected content include, but are not limited to:

- reviewing content where it has been detected as a match with a hash which has not previously been matched with content on the service;
- reviewing content where it has been detected as a match for a hash which previously resulted in a 'false positive' match;

[580] [✂].

- reviewing random samples of content detected as CSAM, prioritising matches where the measured perceptual distance is closer to the threshold used by the technology to determine perceptual similarity; and
- monitoring the frequency of CSAM detected by the technology for spikes in the quantity of content matched and increasing the overall proportion of content reviewed accordingly.

## Benefits and effectiveness

### Benefits

#### CSAM related harm

4.80 CSAM can have a profound and long-lasting impact on children who are sexually abused as well as on the wellbeing of adults and children who unintentionally view this material.

4.81 Beyond the abuse itself, the existence, sharing and viewing of images and videos depicting the abuse can serve as a continual source of trauma for victims and survivors of CSEA. Victims and survivors may experience re-victimisation and can suffer re-traumatisation from everyday triggers, for example heightened sensitivity to photos and cameras.

4.82 Additionally, children themselves may generate content that can be considered CSAM, which can cause them harm.[581]

4.83 CSAM is often viewed intentionally, and the availability of CSAM online creates a permissive environment in which perpetrators can develop and act out their sexual interests. The availability of CSAM can lead to unintentional viewing, which is likely to cause considerable distress.

4.84 While it is difficult to accurately estimate the presence of CSAM online, a number of sources show how widespread it is. The IWF confirmed that it received 275,652 reports containing CSAM, links to CSAM, or advertised CSAM in 2023.[582] Police data also shows that c.107,000 sexual offences against children were recorded by the police across England and Wales in 2022, a 7.6% increase on 2021 and a near quadrupling of the number recorded ten years prior. Police estimates suggest that online CSEA accounts for at least 32% of the recorded total. [583]

4.85 Evidence also suggests the presence of CSAM online is increasing. There have been year-on-year increases in the number of URLs which contain CSAM reported to the IWF, with an 8% increase between 2022 and 2023.[584] As a result, the severe risks posed by CSAM will also increase and, therefore, the removal of this content is essential to ensure the safety and well-being of online users, children and victims of sexual abuse.

4.86 The Register further evidence of CSAM related harm online (see chapter title 'Child Sexual Exploitation and Abuse (CSEA)').

#### Benefits of removing known CSAM online

4.87 We expect that removing known CSAM online will deliver extremely important and wide-reaching benefits.

[581] UK law enforcement refers to this as self-generated indecent imagery (SGII).

[582] IWF, 2024. IWF Annual Report 2023 #behindthescenes. [accessed 8 July 2024].

[583] National Analysis of Police-Recorded Child Sexual Abuse & Exploitation (CSAE) Crimes Report - January 2022 to December 2022 (vkpp.org.uk) [accessed 20 September 2024].

[584] IWF, 2024. IWF Annual Report 2023 #behindthescenes. [accessed 8 July 2024].

- Detecting and removing CSAM can disrupt offending and lead to investigative action against those sharing and viewing such material online. More widespread detection and reporting of illegal content means that perpetrators can be identified, leading to the arrest and conviction of those possessing illegal material and those committing or facilitating child sexual abuse offences. This can in turn prevent perpetrators from committing further child sexual abuse.
- Removing CSAM on online services may result in a reduction of other types of child sexual abuse, such as grooming and contact sexual abuse. Studies indicate a connection between viewing CSAM and committing other child sexual exploitation and abuse (CSEA) offences, including going on to contact children for the purposes of sexual abuse.[585] One study found that 42% of self-reported perpetrators who had viewed CSAM online went on to seek direct contact with a child afterwards.[586]
- Removing CSAM would help reduce both the potential for inadvertent viewing of CSAM and the associated harmful impacts of doing so. Some users may inadvertently view CSAM online as a result of its wide availability. This may be a traumatic experience, provoking feelings of guilt and shame in users. For others, this initial exposure can lead them to search for and intentionally view CSAM.[587]
- Detecting known CSAM points service providers and law enforcement towards communities of perpetrators or locations online where further content is being stored and shared. This can lead to the discovery of unknown CSAM, which can then be reported and hashed for inclusion into the databases to prevent its further circulation. It can also enable the identification of previously-unknown victims and survivors, who can then be safeguarded or protected.
- Removing CSAM on online services can help provide reassurance to victims and survivors that providers are taking proactive measures to address the sharing of content depicting their abuse. Victims and survivors of child sexual abuse are known to experience re-traumatisation as a result of conscious or unconscious reminders that their images are circulating online.[588] Re-traumatisation can involve victims re-experiencing the original trauma.[589] This could be triggered by victims and survivors inadvertently seeing their images or purposely seeking them out in order to report them to online services.

4.88 We acknowledge that hash-matching technologies are not themselves capable of detecting unknown CSAM – that is, CSAM images that are not already included in hash databases. This was raised by [✂] and IJM in response to the November 2023 Consultation, as set out in paragraph 4.56. However, through the detection of known illegal content and identification of offenders by law enforcement authorities, this measure can lead to indirect

[585] For more information on the interrelated nature of CSEA harms, see the Register of Risks chapter titled 'CSEA' (specifically the sections on CSAM, and Grooming).

[586] Insoll, T., Katariina Ovaska, A., Nurmi, J, Aaltonen, M. and Vaaranen-Valkonen, N., 2022. Risk Factors for Child Sexual Abuse Material Users Contacting Children Online: Results of an Anonymous Multilingual Survey on the Dark Web, Journal of Online Trust & Safety, 1 (2). [accessed 22 October 2024]

[587] Wortley, R., Findlater D., Bailey A., Zuhair D., 2024. Accessing child sexual abuse material: Pathways to offending and online behaviour, Child Abuse & Neglect, 154. [accessed 06 November 2024].

[588] The CSAM survivors' group Phoenix 11 responded to the November 2023 Consultation by stating that CSAM survivors "cannot heal" while images depicting "the most horrific moments of [their] lives" is circulating freely online. Source: Letter to Ofcom Online Safety Team • Phoenix 11 (thephoenix11.com). [accessed 18 October 2024].

[589] Fisher, C., Goldsmith, A., Hurcombe, R., and Soares, C. (IICSA Research Team), The impacts of child sexual abuse: A rapid evidence assessment. (July 2017), [accessed 30 November 2024].

benefits that include the detection of unknown CSAM and reduced contact CSEA. Even though these benefits arise indirectly, they can be very substantial (as illustrated in Annex 5).

### Effectiveness

4.89 The use of perceptual hash matching is a well-established industry practice and has resulted in the detection and removal of millions of items of CSAM.[590] For example, an analysis of NCMEC data found that a major contributor to the exponential growth in reports made to its CyberTipline since 2009 is the increased use of proactive automated detection tooling (such as perceptual hash matching).[591] Between 2010 and 2020, the number of reports to the CyberTipline increased from just over 10,000 per year to over 21 million per year.[592] It is understood that the vast majority of reports to NCMEC (specifically, reports of CSAM) are generated by perceptual hash matching technology. While these statistics are somewhat limited in that they relate only to known images reported to a US reporting body, they provide a strong indication that hash-matching is effective in detecting known CSAM.

4.90 The effectiveness of the measure depends in large part on the quality of the set of hashes used for hash-matching, including whether the material from which the hashes are derived has been accurately assessed to be CSAM, the range of material included in the set of hashes, and the number of hashes included.[593] The measure includes a number of requirements designed to ensure:

- the set of hashes used by the service provider reflect the range of CSAM that is illegal under criminal law in the UK;
- the set of hashes includes hashes sourced from one or more organisations with expertise in the identification of CSAM;
- the sourced database is regularly updated with newly discovered content; and
- there are governance arrangements in place to ensure that CSAM is included in the hash database correctly, and to allow CSAM hashes to be reviewed and removed swiftly if found to be incorrect.

4.91 We set out requirements for a hash database to be appropriate to use for the measure in paragraph 4.68 to 4.72. We consider that these provide sufficient assurance that the database(s) used will support an effective use of hash-matching and gives guidance to support service providers in sourcing an appropriate database. This guidance also responds to stakeholder feedback, as detailed in paragraph 4.24, requesting parameters of an acceptable hash database.[594]

4.92 The accuracy and quality of hashes also relies on the clarity of the definition for the type of content to be detected and removed. Stakeholders raised that this is an important consideration when selecting a database to use. We’ve made clear that the set of hashes to be used should reflect the range of child sexual abuse material that is illegal in the UK. We

[590] For example, in 2023 NCMEC labelled (and subsequently hashed) over 10.6 million files. NCMEC, CyberTipline Data [accessed 22 October 2024].

[591] Burstein, E., Clarke, E., DeLaune, M., Elifff, D.M., Hsu, N., Olson, L., Shehan, J., Thakur, M., Thomas, K. and Bright, T., 2019, May. Rethinking the detection of child sexual abuse imagery on the internet. In The world wide web conference (pp. 2601-2607).

[592] Burstein, E., et al. 2019.

[593] The importance of the accuracy and quality of hashes was reiterated by stakeholders in response to the November 2023 Consultation (refer to paragraph 4.25).

[594] [✂].

have explained above that (in particular) this includes indecent images of children that do not depict sexual activity (known as ‘Category C imagery)’, and non-photographic CSAM. More information is set out in the Illegal Content Judgements Guidance. [✂].[595] We have also made changes to the measure (set out in paragraph 4.71) to ensure service providers can use hash databases from “one or more” organisations with expertise in the identification of CSAM.

4.93 These changes mean, for example, that it would not be sufficient for a service provider to use only a hash database that is limited to illegal child sexual abuse material in another jurisdiction where that jurisdiction uses a materially narrower definition of CSAM. The service provider would need to ensure that it also used other hashes to address this (for example, by also using a hash database of material assessed to be CSAM in accordance with UK law).

4.94 These changes to the measure account for legal definitions of CSAM in other jurisdictions which some hash sets are based on. We consider the changes are likely to make the measure more effective, as a greater volume of CSAM will be captured compared to our original proposal.

4.95 We are aware that some service providers may use other hash sets to detect content which is prohibited under their terms of service, such as content that is associated with child exploitation. We recognise these service providers’ efforts to protect children, and this measure does not prevent service providers from adopting such safety measures themselves. If, however, a service provider wishes to use a broader database (i.e. one that includes hashes of material other than CSAM) for the purposes of this measure, it will need to ensure that it complies with the measure’s provision at least in relation to hashes of CSAM content in the database.

4.96 The use of a broader database could increase the risk of incorrect removals of content. Our codes measures about reporting and complaints enable users to make complaints if their content is taken down on the basis that it is illegal content. They help to mitigate the impact of content being incorrectly identified as CSAM. As mentioned above, we are aware that some service providers may use hash sets which extend beyond the definition of illegal content under the Act. Where service providers seek to use such hash sets for the purposes of this measure, we would expect them to enable users to make complaints if their content is detected as CSAM when implementing this measure and taken down as a result of using hash-matching technology.

4.97 The measure does not require the use of specific databases to facilitate flexibility and autonomy of service providers in their content moderation practices. This also helps address accessibility of hash databases for service providers by not prescribing the use of a specific database. The need for flexibility in our approach was also raised by stakeholders in response to the November 2023 Consultation (refer to paragraph 4.23). Comparatively, we received feedback from an organisation providing hash databases, proposing that the measure recommends specific databases (refer to paragraph 4.27). This measure allows service providers the flexibility to choose a third-party database of CSAM hashes. There are several reasons why we do not think it is appropriate to prescribe the use of a specific database. The reasons include:

[595] [✂].

- We do not have access to information on every active organisation providing hash databases that would allow us to recommend one organisation over others who may be capable of performing the same function.
- We want to ensure that the measure is future-proof. Recommending a specific organisation could undermine this aim as the availability of organisations providing hash databases is subject to change over time. This approach would require extensive oversight to ensure quality and effectiveness.
- Recommending a centralised organisation would limit the ability of service providers to choose a database for their services.

4.98 Ultimately, we have decided not to prescribe the use of specific hash databases, and service providers can choose any appropriate hash database which meets the requirements set out in the measure. We understand, as set out in the stakeholder feedback (refer to paragraph 4.27), that there are also several third-party database providers that can be accessed by service providers in order to successfully implement this measure. This provides them with greater flexibility.

4.99 Our analysis suggests that the measure will be highly effective at detecting known CSAM. It will therefore materially reduce the circulation of CSAM compared to a counterfactual in which service providers did not apply the measure. Given the incredibly severe harm that the circulation of CSAM causes, the measure will have very material benefits.

#### Other approaches

4.100 As described in the 'Summary of stakeholder responses' some stakeholders recommended expanding the kinds of harms the measure applies to, including to gender-based harms or adult intimate image abuse or terrorism. Some other stakeholders commented that the measure only facilitates detection of known CSAM and suggested alternative systems and processes to identify new CSAM.

4.101 We agree that a range of interventions are needed to tackle the harm of CSA. Our approach will come in stages and should be looked at holistically. We are initially combatting abuse and the generation of CSAM upstream with our safety default and supportive information measures (see chapter 8 of this Volume: 'U2U settings, functionalities, and user support'), and combatting the onward circulation of CSAM and the associated re-traumatisation downstream with our hash-matching and URL detection measures. There is more to do to address first generation CSAM. We are taking steps to address this type of content by exploring future measures that recommend more sophisticated automated moderation tooling, as it becomes more widely available. We intend to consult on additional measures in spring 2025. This will include work we announced earlier this year, to consult on how automated tools can be used to proactively detect illegal content and the content most harmful to children, going beyond the automated detection measures we are recommending in this chapter.

4.102 In summary, the circulation of CSAM causes substantial harm. Whilst many of the largest service providers are already hash matching, we expect our measure to significantly expand the range of providers that are hash matching for known CSAM – in particular, we think that the measure will significantly expand the number of high-risk services that use hash matching. We therefore consider that the decision we are taking will result in a sizable increase in the volume of CSAM that is detected and removed. Given the harm caused by CSAM, this will have very significant benefits.

## Costs

4.103 This section summarises our assessment of the costs associated with implementing the measure. Service providers are likely to incur both one-off set-up costs and ongoing maintenance and operating costs. These 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.

4.104 Where possible, we have updated the relevant data underlying our cost estimates for the latest year available (for example, wage data), as well as for additional insights obtained (for example, on the annual cost of software, hardware, and data), which has resulted in some changes to our cost estimates as set out in the November 2023 Consultation.

4.105 The costs of a service provider implementing the measure will depend on numerous factors, including a service's number of users and a service's risk of image-based CSAM. As we are recommending the measure to a broad range of services, so too do we expect a broad range of costs. We have quantified costs by creating 'hypothetical' services, defined in terms of their number of users and risk of image-based CSAM, that would be in scope of the measure. For example, we have modelled a hypothetical service with 700,000 monthly UK users that is high-risk for image-based CSAM.

4.106 See Annex 5 for more detail on how we have quantified these costs for the hypothetical services; the feedback received from stakeholders on these costs; and our assessment of costs in light of this feedback and changes to the scope of our measure.

### One-off costs

4.107 The one-off cost of building a hash-matching system will primarily be labour costs. Software engineers will be required to build the system, supported by a range of other professionals such as product managers, analysts, and lawyers. The technological solution used to integrate hash-matching into a service will affect these development costs.[596] Our understanding is that it is cheaper for service providers to integrate hash-matching via a third-party API to access hash-matching functionalities provided by a third party than by building an in-house hash-matching system. The technical complexity of the service also affects these costs – integration will be more challenging for services with a larger number of relevant functionalities and larger technology stacks of such functionalities, which can increase the resource required to implement a new technology.

4.108 Through our engagement with industry experts and stakeholders, we understand that building a CSAM hash-matching system may take around two to 18 months of full-time work by a software engineer. We have also included equivalent time for other professional occupation staff. We estimate one-off set-up costs to be between £17,000 and £339,000 depending on the size and complexity of the service (as these factors would impact the time

[596] In response to our November 2023 Consultation, one stakeholder, INVIVIA pointed out that the integration of hash matching can be particularly costly for smaller service providers. INVIVIA response to November 2023 Consultation, pp.15-16. We consider these comments further in Annex 5.

required to build the system).[597] In general, we would expect the costs smaller services incur to be towards the bottom end of this range.

### Ongoing maintenance costs

4.109 Ongoing costs include the labour costs of maintaining the hash-matching system. Activities include applying updates, reviewing the technology's performance and adjusting parameters (at least every six months), ingesting new hash lists, and integrating with new functionalities.[598] Consistent with our standard assumption for the ongoing costs of system changes, we assume that annual maintenance costs are 25% of the initial set-up costs.[599] As with the cost of building a hash-matching system, the cost of maintaining the system is likely to scale with the technical complexity of the service. We expect larger service providers to generally be more complex and require more bespoke systems to be built and maintained. The annual cost of maintaining a hash-matching system is estimated to range from £4,000 to £85,000 depending on the size and complexity of a service.[600] Once again, we would expect the costs incurred by providers of small services to lie towards the bottom of this range.

### Ongoing software, hardware, and data costs

4.110 Ongoing costs will also include the annual cost of software, hardware, and data. These costs will generally be larger for service providers with a large user base because it is common practice for non-governmental organisations in the hash-matching ecosystem to charge based on the capacity of the service provider to pay for their product.[601] Providers of larger services, especially those with more complex product portfolios, may also more often opt for in-house solutions that require multiple hash lists and software products. Although lower-cost solutions are available, we estimate that the annual cost of software, hardware, and data could start at £1,000 for small services,[602] and theoretically go up to £1 million for a service that reaches the entire UK population.[603]

[597] We expect the cost for most services to fall within the estimated ranges, but we are aware that there may be exceptions on either side of this range, because some services in scope of our measure will be smaller or larger than those that have been modelled.

[598] This is also consistent with the cost activities that were identified by a few stakeholders ([✂], INVIVIA, UK Interactive Entertainment (Ukie)). Name withheld 5 response to November 2023 Consultation, p.11; INVIVIA response to November 2023 Consultation, pp.15-16; Ukie response to November 2023 Illegal Harms Consultation, p.20.

[599] Standard assumptions on costs are detailed in Annex 5.

[600] Again, we expect the cost for most services to fall within the estimated ranges, but we are aware that there may be exceptions on either side of this range, because some services in scope of our measure will be smaller or larger than those that have been modelled.

[601] For example, IWF and Thorn charge services based on factors such as company size, capacity to pay, number of API queries, etc.

[602] We have updated this figure since the November 2023 Consultation to reflect additional insights obtained. For example, we are aware that IWF's membership fees, which vary based on industry sector and company size, can start at £1,000 per year. These fees would include access to a hash list as well as hash-matching technology such as PhotoDNA. IWF, 2024. Membership fees. [accessed 4 November 2024]; and IWF, 2024. Image hash list. [accessed 4 November 2024]. We are also aware that these costs may be lower where a service is deemed by a relevant NGO provider to have less ability to pay, and that there may be other low cost or free options available in the market.

[603] These assumptions are based on our own expertise and industry experts, which includes consideration for the membership fees that apply to service providers to access hash lists and the price of all-in-one software solutions. A few stakeholders pointed towards the costs associated with accessing hash lists for smaller service

### Ongoing content moderation and reporting costs

4.111 Another ongoing cost is the labour costs related to reviewing, moderating, and reporting CSAM.[604] In general, these costs will scale with the amount of content that is matched by the hash-matching system. The number of matches is mainly determined by the risk of CSAM being present on the service. A higher risk of CSAM increases the likelihood of known CSAM being detected, which may need to be subject to human review (including in the case of relevant user complaints and appeals) and reported to relevant authorities. The number of matches is also determined by the number of false positives, which depends on factors like the size of a service and amount of content on a service. That said, we expect that service providers will be able to control (to an extent) the number of false positives when configuring their technology, allowing them to strike an appropriate balance between precision and recall.

4.112 As detailed in Annex 5, we have assumed that moderation costs depend on the size of a service and the risk of CSAM being present on a service. Based on our own expertise and on our engagement with industry experts and stakeholders, we expect that service providers with a small user base and a relatively low risk of CSAM will not require a full-time moderator, whereas high-reach and/or high-risk service providers will likely employ a team of moderators. We estimate that a provider of a high-risk service with 700,000 monthly UK users could spend between £18,000 to £55,000 on moderators per year. This cost could be larger for services with more users or at higher risk of CSAM. As moderation costs are broadly correlated with the amount of CSAM on a service, we would expect benefits to increase in line with such costs.

4.113 In response to our November 2023 Consultation, some stakeholders highlighted the importance of human review in determining whether content detected through hash matching is CSAM.[605] We agree with these comments and consider human review to be an important part of our overall measure. Several stakeholders commented that the costs associated with reviewing, moderating, and reporting of CSAM could be particularly burdensome for providers of small and medium sized services.[606] Considering these comments, we engaged further with a number of stakeholders to understand the likely human moderation costs that may be incurred by providers of smaller services as a result of this measure (see Annex 5). In summary, although we acknowledge the variation in the costs likely to be incurred across services, we remain of the view that our estimates for the costs associated with human content moderation are reasonable.

providers, and the burden this can place on those providers. One stakeholder also commented that we did not account for the membership fees needed to access hash lists. We consider these comments in further detail in Annex 5.

[604] We anticipate many services will develop or access some form of automated reporting system, meaning reviewers will be able to report matches in a streamlined manner, reducing the time and cost spent on reporting individual matches.

[605] Name withheld 5 response to November 2023 Consultation, pp.10, 15; Microsoft response to November 2023 Consultation, p.13; [✂].

[606] INVIVIA response to November 2023 Consultation, pp.14-15; [✂]; Integrity Institute response to November 2023 Consultation, p.12.

4.114 We acknowledge that there may be other costs to the service provider,[607] as well as costs (both monetary and non-monetary) to other organisations and individuals. For example, potential costs associated with the measure to the broader hash-matching ecosystem were highlighted by some stakeholders in their responses to the November 2023 Consultation responses (see paragraph 4.35). We also recognise that service providers or third-party organisations (such as hash database providers) could incur additional costs for implementing measures to secure hash databases (see paragraph 4.72 for examples of the possible security measures).

### Total costs

4.115 In response to the November 2023 Consultation, several stakeholders noted that some of the one-off and ongoing cost components of implementing this measure could be particularly material for providers of smaller services.[608] This is why we have focused our quantitative analysis on hypothetical services that vary in terms of their risk for image-based CSAM, as well as their user numbers. We discuss this in more detail in Annex 5.

4.116 We acknowledge that providers of smaller services may experience additional barriers to undertaking hash-matching, such as:

- a lack of specialist policy knowledge about CSAM;
- a lack of pre-existing engineering expertise to integrate perceptual hash-matching across their products;
- a lack of human resources to dedicate towards the reviewing, moderating, and reporting of CSAM.

4.117 While we acknowledge these potential barriers, we are aware that it is not uncommon for some non-governmental organisations (NGOs) to work with providers of smaller services so that they can overcome these types of barriers. We are also aware that some smaller service providers are already using perceptual hash-matching to identify known CSAM, which suggests that these barriers are surmountable for service providers with small user bases. The type and size of service providers that deploy hash matching is varied and includes many different types as well as different sizes of service.

4.118 We estimate the total costs associated with implementing and maintaining a hash-matching system for two hypothetical services: the smallest service with a medium risk for CSAM that we will be recommending hash matching to (seven million monthly UK users), and a much smaller service (70,000 monthly UK users) with a high risk for CSAM.[609] We estimate that the one-off costs of implementing this measure could range from £17,000 for the service that reaches 70,000 monthly UK users to £339,000 for the service that reaches seven

[607] For service providers, there are likely to be further costs associated with implementing and having policies to take down detected content that is CSAM. For example, there are likely to be costs associated with putting in place protocols for those dealing with the hash-matching system to ensure that the relevant material is handled securely (for example, ensuring that only named individuals are permitted access to and involvement with the implementation, testing, and review of the system).

[608] INVIVIA response to November 2023 Consultation, pp.14-15; Microsoft response to November 2023 Consultation, p.13; [✂]; Integrity Institute response to November 2023 Consultation, p.12; Scottish Government response to November 2023 Consultation, p.7.

[609] Annex 5 provides a more complete analysis: it includes an additional hypothetical service (one with a high risk for CSAM and that reaches 700,000 monthly UK users) so as to align with the three types of service providers that we are recommending hash matching to; it projects the costs into the future and applies additional sensitivities; and it gives a sense of how costs stack up against the benefits that we expect this measure to achieve.

million monthly UK users. We also estimate that the ongoing annual costs could range from £8,000 for the service that reaches 70,000 monthly UK users to £224,000 for the service that reaches seven million monthly UK users.[610] These estimates are based on the costs discussed above and the assumptions further outlined in Annex 5.

4.119 We recognise that there will be services in scope of the measure that will have fewer than 70,000 monthly UK users and services that will have more than 7 million monthly UK users. Even for services whose number of users are within this range, we recognise that some service providers may incur higher costs. For example, we estimate that a service that reaches seven million monthly UK users and that has a high risk of CSAM could incur an annual ongoing cost of around £732,000. Costs above the ranges presented in paragraph 4.119 would likely be explained by a higher risk of harm, and so higher costs would be matched by higher benefits of adopting the measure. Further detail on our estimates of total costs for different sizes and kinds of services, and the updates made since the November 2023 Consultation, can be found in Annex 5.

## Risks

4.120 This section discusses risks of:

- incorrect detection of content as CSAM;
- security compromises; and
- potential biases in hash databases.

4.121 Our view is that each of these risks can be sufficiently mitigated where hash matching is implemented in accordance with our measure and the safeguards we have set out. In particular, we consider that the measure should result in relatively small volumes of content being incorrectly detected as CSAM. We discuss the impacts of this further in the "rights impacts" section.

### Incorrect detection of content as CSAM

4.122 There are two main reasons why content could be incorrectly detected as CSAM: first, if a hash had been incorrectly included in a database used for hash matching; and second, where content detected by the hash matching technology as a match for CSAM is not CSAM (known as a "false positive").[611]

#### The accuracy of hash databases

4.123 We recognise there is a possible risk of content being incorrectly identified and added to a hash database. As described in the 'effectiveness' section, our measure provides for hashes to be sourced from organisations with expertise in identifying CSAM. It sets out

[610] In our November 2023 Consultation, we estimated the total costs for the smallest service providers that would have been in scope of our proposed measure. We estimated that the one-off costs of implementing this measure could range from £16,000 for a service that reaches 70,000 monthly UK users, to £319,000 for a service that reaches seven million monthly UK users. We also estimated that the ongoing annual costs could range from £31,000 annually for a service that reaches 70,000 monthly UK users to £254,000 for a service that reaches seven million monthly UK users. We recognised that the costs were likely to vary further with the risk of CSAM being present on services. For example, we estimated that a service with seven million monthly UK users and at high risk of CSAM could incur an ongoing annual cost of £820,000. Such a service may encounter a higher volume of CSAM and therefore require more moderators to review and report the content.

[611] BILETA response to November 2023 Consultation, p.12; Name withheld 5 response to November 2023 Consultation, p.10; Global Partners Digital response to November 2023 Consultation, p.14; Integrity Institute response to November 2023 Consultation, p.13; INVIVIA response to November 2023 Consultation, p.14; Proton response to November 2023 Consultation, p.6; [✂].

requirements for arrangements to secure (so far as possible) that CSAM is correctly identified before being added to the database and to review cases where hashes are suspected to have been added incorrectly.

4.124 NGOs providing these databases strive to operate to high standards. While assessing whether material is CSAM can be difficult, the available evidence points to very high levels of accuracy in their decision-making.[612] [613] For example, in 2023 an independent audit of NCMEC's database of CSAM found that 99.99% of images and videos met the US federal definition of "child pornography".[614] [615] NGOs have also published reports (or provided information to us) describing the robust controls in place to ensure the accuracy of their databases.[616]

4.125 As described in the "how this measure works" section, the measure also allows for use of CSAM hashes not sourced from NGOs (such as of CSAM identified by the service's content moderation function). Safeguards have been included in the measure to reduce the risk of this content being incorrectly hashed as CSAM. While the controls implemented by some service providers might not be as robust as those operated by NGOs, they can mitigate the risk further through their approach to human review (see paragraph 4.74).[617]

4.126 We consider the elements set out in this measure to ensure accuracy of hash databases are appropriate to mitigate risk of content being incorrectly added to a database.

### False positives

4.127 False positives occur where two visually distinct images are treated as a match.[618] The risk can be controlled through configuring the technical parameters used to determine if there is a match and through systems and processes used to detect false positives (in particular, human review).[619]

4.128 Configuring the technical parameters involves striking a balance between the technology's recall and precision: for example, setting parameters that treat two hash values that are

[612] [✂] commented on the minimal level of errors that had been [✂]. [✂].

[613] Assessing age can be particularly difficult: see for instance IWF, Internet Watch Foundation Inspection Report 2022. (30 January 2023), p.5 [accessed 18 October 2024].

[614] NCMEC, Attestation report issued upon completion of Concentrix's audit of NCMEC's hash list. (12 April 2024) [accessed 18 October 2024]. This database acts as the source file for the perceptual hashes made available to service providers. The audit identified 59 'exceptions' among the 538,922 visually distinct images and videos on NCMEC's CSAM Hash List, which were stated to have been subsequently removed from that list. The report does not confirm whether hashes contributed by NCMEC to the Non-Governmental Apparent Child Pornography Hash-Sharing Initiative of content identified as visually similar to those exceptions were also removed.

[615] [✂] also commented on the minimal level of errors that had been [✂].

[616] These controls include training of analysts, assessment by multiple analysts or quality assurance, and processes to review any suspected errors (such as complaints procedures). For example: The Internet Watch Foundation (IWF), Information available on its website about its Image Hash List includes a section on "enhanced quality checks" where it explains: "Each child sexual abuse image or video is independently assessed by at least two people. These enhanced quality checks mean every image or video is categorised correctly before it is added to the hash list. [Accessed 14 November 2024].

[617] For instance, service providers can review matches for hashes where they have not been matched with content before.

[618] As there are a finite number of possible hash values, it is also theoretically possible for two distinct images to have the same hash value (resulting in a 'hash collision').

[619] See also paragraph 4.74.

further apart as a match should improve recall but increase the risk of false positives.[620] Our measure expects service providers to strike an appropriate balance, taking specified matters into account, and to review this at least every six months. As described in the "how the measure works" section, the measure also sets out that providers should ensure that an appropriate proportion of detected content is reviewed by human moderators.

4.129 There is limited direct evidence available about the level of false positives resulting from perceptual hash matching for CSAM.[621] [622] However, our overall assessment is that if hash matching is implemented in accordance with the safeguards we have set out, it should have a very high level of accuracy and give rise to relatively small numbers of false positives.[623]

4.130 Meta's reports to the European Commission reflect the outcomes of its use of hash matching for CSAM and human review of some detected content, and therefore provide a useful guide for the levels of content that might be inaccurately actioned as CSAM. In 2022 and 2023, it restored around 3,600 and 11,600 items of content respectively after identifying false positives following user appeals.[624] This is to be seen in the context of actioning around 6.6 million and 3.6 million items of content based on detection by its media-matching technology.

4.131 The false positive rates on some other services could be significantly higher. This is in particular because of the "false positive paradox":[625] if the prevalence of CSAM on a service is very low, then even highly accurate technology could result in a considerable proportion of false positives in detected content. In such cases, we anticipate that service providers will adjust the technology's configuration as needed to ensure they are able to address this (for instance, by reviewing all or a high proportion of matches).[626] Our measure sets out a 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/processes). Service providers should take this principle into account when deciding

[620] Recall is the probability of detecting (in this case) CSAM: i.e. the proportion of all positives that were classified as positive. Precision is the proportion of positive classifications that are actually positive.

[621] Responses to the November 2023 Consultation provided little quantitative evidence. Some commented that false positives were rare, while others emphasised the importance of human review (implying a level of false positives).

[622] A recent analysis of PhotoDNA (a hash matching technology widely used to detect CSAM) found that using a distance threshold that provided a useful level of robustness to image transformations resulted in a false positive rate of 0.3% in test conditions. Steinebach, M. An analysis of PhotoDNA. (2023), p.5 [accessed 18 October 2024].

[623] We recognise that the level of false positives could initially be higher where hash matching is first implemented on a service, as technical parameters and human review policies are adjusted.

[624] Meta, EU CSAM derogation report (30 January 2023) and EU CSAM derogation report (31 January 2024) [accessed 18 October 2024]. C3P described the earlier report as providing "a real-world baseline for accuracy rates on a very high volume provider": C3P response to November 2023 Consultation, p.19. These reports relate to message threads on Messenger and Instagram which involved an EU user. We recognise that this data reflects errors identified through appeals and that (among other things) not all false positives will be appealed. However, while a much higher proportion of content actioned was appealed in 2023 than in 2022 (7.07% vs. 0.44%), the proportion of successful appeals fell by around two-thirds, suggesting that it would be unreasonable to assume that the proportion of successful appeals can be extrapolated to all detected content.

[625] Hunt T, Williams B, Howard D. The "false positive paradox" and the risks of testing asymptomatic people for COVID-19. Authorea Preprints; 2021. doi: 10.22541/au.162462792.21745309/v1. PPR:PPR361967.

[626] In cases where a high proportion of false positives are identified, we anticipate services will address this by baselining their performance metrics and fine-tuning the parameters of their systems by adjusting selected metrics to fit their individual workflows and deployment environments, enabling them to review all, or a high proportion, of detected content.

their policies for review and ensure that potential false positives are dealt with appropriately.

4.132 Provided they do so and provided they adhere to all other safeguards built into our measure, we consider that the risk of false positives should be manageable.

### Security compromises

4.133 A number of stakeholders commented on risk of security compromises of hash databases or hash matching technology. This could, for example, enable hashes to be added or removed from a database without authorisation.

4.134 As described in the 'how this measure works' section, our measure sets out a requirement for organisations from which hash databases are sourced to secure those databases against insider and outsider threats. NGOs providing these databases should already have these controls in place. The measure also sets out that service providers should ensure an appropriate policy is in place, and that security measures are taken in accordance with that policy, to secure any hashes held for the purposes of the measure. Securing hashes sourced from an NGO will usually be a contractual requirement. We describe good practice for mitigating security risks at paragraph 4.72. We therefore consider that the safeguards we have put in place are sufficient to mitigate these risks appropriately.

4.135 If (as we expect) there is a significant expansion in the number of services using hash matching, we also recognise that the risk of adversarial attacks on hash matching technology could increase. As with hash databases, we expect those providing access to hash matching functions to impose contractual controls to limit this risk. We have also considered whether these concerns could result in barriers to service providers being able to implement the measure. We note that small services (which may be less able to implement robust security controls) can access hash matching through API-based[627] solutions where these risks do not arise.

### Potential biases in hash databases

4.136 A few stakeholders commented, in principle, on risks that could arise if hash databases contained biases. While no specific evidence of bias was provided, it was noted that how CSAM is identified and assessed could create biases that underrepresent particular kinds of victims and survivors, with possible unintended consequences.

4.137 In response to this feedback, we have included an additional requirement in the measure. Service providers should ensure that the arrangements operated by the organisation(s) from which hash databases are sourced for identifying or assessing suspected CSAM do not plainly discriminate on the basis of protected characteristics (such as sex or race).[628]

4.138 We consider this should safeguard against systematic biases (for example, if an organisation were to exclude CSAM relating to certain kinds of victim as a matter of policy). We have made similar changes to our other automated moderation measures.

[627] [✂].

[628] The "protected characteristics" (as specified in Part 2 of the Equality Act 2010) are age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex and sexual orientation.

### Conclusion

4.139 Having reviewed the evidence available to us, we consider that the risks associated with perceptual hashing can be managed, and it is appropriate to proceed with recommending this measure to address the serious harm from CSAM.

4.140 As discussed in this section, risk levels can be reduced and mitigated through the practices of both service providers implementing hash matching and organisations which compile and maintain hash databases. We have therefore designed the measure with a number of safeguards detailed in this chapter and adjusted some of the safeguards following stakeholder feedback to our consultation.

## Rights impact

4.141 This section considers the measure's impacts on users' rights under Articles 8 and 10 of the European Convention on Human Rights ('ECHR').

4.142 As explained in 'Introduction, our duties, and navigating the Statement', as well as in chapter 14 of this Volume: 'Statutory tests', Article 10 of the ECHR sets out the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 8 of the ECHR sets out the right to respect for individuals' private and family life.

4.143 In essence, restrictions on those rights must be necessary and proportionate – that is, the measure's contribution to its objective must outweigh its adverse impacts.

4.144 Our assessment of those adverse impacts is therefore to be balanced against the measure's contribution to removing CSAM. Parliament has legislated for CSAM to be 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, relating to each of the prevention of crime, the protection of health and morals, and the protection of the rights of others.

4.145 Detecting, removing, and reporting CSAM acts to prevent crime by deterring users from posting such content and by preventing other users from accessing it. It also acts to protect public morals, including by preventing users from inadvertently encountering CSAM. The 'Benefits and effectiveness' section discusses the measure's effectiveness in detecting and removing CSAM and the associated benefits in more detail.

4.146 Removing CSAM is vital for protecting the right to privacy of victims of child sexual abuse and also helps to protect their personal data. C3P's response to the November 2023 Consultation emphasised that "allowing CSAM to be uploaded and shared is a massive violation of the privacy of the person depicted".[629] The 'Benefits and effectiveness' section describes the ongoing harm to victims caused by knowing that material depicting their abuse continues to circulate (or in some cases themselves viewing that material), or from being identified by persons who have viewed that material.

4.147 Measures that restrict the dissemination of CSAM online may also reduce levels of child sexual abuse, both by reducing contact offending associated with viewing CSAM and by enabling offenders involved in contact offending to be identified (through reporting of

[629] C3P response to November 2023 Consultation, p.18.

CSAM to law enforcement authorities).[630] Annex 5 gives a sense of the number of children per year that may be safeguarded from contact child sexual abuse due to hash-matching. 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 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 ECHR.

### Freedom of expression

4.148 An interference with the right to freedom of expression must be prescribed by law, pursue a legitimate aim, be proportionate to the legitimate aim and correspond to a pressing social need.

4.149 Interference with users' freedom of expression (including rights to receive information) arises where content is wrongly taken down by the systems and processes implemented in accordance with the measure.[631] [632]

4.150 As a preliminary point, then, it is important to reiterate that the measure only applies in relation to content communicated publicly (consistent with the constraints on Ofcom's power to include 'proactive technology' measures in Code of Practice).

4.151 The 'Risks' section explains that the risk of incorrect detection of content as CSAM the set of hashes used, the technical configuration of the hash matching technology, and other systems and processes used to detect false positives (including human review).

4.152 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.

a) It sets out that arrangements should be in place to (a) secure that CSAM is correctly identified before hashes of that material are added to the set of hashes used, and (b) review any cases where material is suspected to have been incorrectly identified as CSAM (and remove those hashes where appropriate); as well as providing that the hashes in use should include hashes sourced from one or more persons with expertise in the identification of CSAM. It also sets out that the service provider should regularly obtain the latest version of any external databases for use when carrying out hash-matching.

[630] For a detailed analysis of the links between CSAM and further CSEA offences, see 'How CSAM and grooming offending overlap' in the Register of Risks chapter titled 'CSEA'.

[631] In principle, adverse impacts on freedom of expression 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 that the Act seeks to protect, such as content of democratic importance or journalistic content. However, the measure is unlikely to have a systematic impact on these kinds of content – for instance, such content would be unlikely to be particularly vulnerable to being incorrectly detected as a false positive. The measure is also unlikely to result in the disclosure of journalistic sources to law enforcement authorities, given that it concerns only images and videos and applies only to content communicated publicly.

[632] In principle, there could also be a more general 'chilling effect' if some users were to avoid using services which had implemented the measure due to concerns about adverse impacts on users' rights. However, many UK users already use services which have implemented perceptual hash-matching and are under obligations in US law to report CSAM to NCMEC, which then passes reports on to relevant national authorities including the NCA. Any such effect therefore seems unlikely to be significant.

b) It sets out that both the service provider and any person from whom hashes of CSAM are sourced should have security measures in place to secure the integrity of the data (as set out in more detail in paragraph 4.71 and 4.72).

c) It sets out that the service provider should configure the technology so as to strike an appropriate balance between precision and recall, taking into account specified matters including the risk of harm relating to CSAM on the service and the proportion of false positives detected, and review this at least every six months. It further provides for the service provider to keep a written record of how it has struck this balance.

4.153 These safeguards reflect good practice and should ensure that the implementation of perceptual hash-matching limits the impact on users' freedom of expression.

4.154 The measure also sets out processes designed to further reduce the impacts of false positives detected by the technology:

a) It sets out that service providers should put in place a policy for review of detected content which secures that human moderators review an appropriate proportion of content detected as CSAM. What is appropriate should be decided taking into account specified principles (including the degree of accuracy achieved by the technology). It also provides that the service provider should ensure a written record is made of its policy.

b) It specifies other Codes measures as safeguards for users' freedom of expression, in particular those relating to content moderation, enabling users to complain if their content has been taken down on the basis that it is illegal content, and ensuring that the terms of service give information about the proactive technology used and the policies and processes for complaints.

4.155 These other measures help to safeguard users' freedom of expression in a number of different ways, including ensuring that (where those other measures apply to the service in question) the service provider sets internal content policies and provides training and materials to content moderators, which would support them in determining whether detected content has been accurately identified, and in providing a level of transparency for users about the technology used and how to make a complaint.

4.156 As discussed in the 'Risks' section, however, we still expect there to be relatively small numbers of false positives which are not identified through proactive human moderation. The number of such cases and consequent impact on freedom of expression will depend on the service in question (i.e. the kind of content communicated publicly on the service) and the precise way in which the service provider implements the measure.

4.157 In this respect, the measure provides service providers with flexibility (for instance, as to the proportion of detected content that should be reviewed by human moderators). We recognise that this could lead to a significant variation in the impact on users' freedom of expression between services, including implementations that have more substantial adverse impacts.[633]

4.158 In practice, service providers have reputational and commercial incentives that should encourage them to strike a reasonable balance in their implementation of the measure. While service providers have incentives to minimise the amount of CSAM on their platform

[633] We also recognise that service providers will adjust the technical configuration of their hash-matching system and other processes over time, in light of their experience. This may mean that the impact on users' rights is more significant in the earlier stages of deployment.

and to limit the costs of human moderation, incentives also exist to limit the amount of content that is wrongly taken down and reduce the costs of handling user complaints. We therefore consider that evidence from service providers' existing use of hash-matching (which is influenced by these incentives) provides a useful guide to the adverse impacts that could be expected.

4.159 Where a service provider also uses hash-matching to detect other content (such as content that is prohibited by its terms of service), it should ensure that the safeguards in this measure are applied at least in relation to hash-matching for CSAM. For instance, if the provider were unable to distinguish whether content had been detected as a match for CSAM or for other hashed material, human moderators would need to review an appropriate proportion of all detected content to ensure the safeguards applied to content detected as CSAM.

4.160 Interference with users' freedom of expression may also arise where service providers take action against those users (such as banning the account) because the user has been detected as sharing (or viewing) CSAM. We intend to propose including a code measure in future about the action service providers should take against users engaged in such illegal activity. However, at this stage, the Codes do not include such a measure and the action to be taken would be a matter for the service provider. Such action could have a more significant impact than take-down of content, and it would be important for service providers to have regard to that impact when deciding on their safety policies.

4.161 However, the safeguards included in this measure to protect users' freedom of expression would by extension limit the risk that action is taken against users on the basis of false positives. We also note that service providers are required to enable users to make complaints if the provider has given a warning to the user, suspended or banned the user from the service, or in any other way restricted the user's ability to use the service, as a result of content shared by the user which the provider considers to be illegal content.

4.162 Overall, while we acknowledge the measure involves interference with users' right to freedom of expression where content is incorrectly detected as CSAM, we consider that interference to be limited with the safeguards for freedom of expression we have in place and proportionate to the measure's aim of reducing the prevalence and dissemination of CSAM.

### Privacy

#### Data protection

4.163 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, either because the user 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. 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.[634]

4.164 In designing Code measures, we have borne in mind the importance of ensuring service providers comply with both online safety and data protection rules. Service providers will

[634] ICO response to November 2023 Consultation, p.12. See also ICO, 2022: Overview of Data Protection Harms and the ICO's Taxonomy. [accessed 18 October 2024].

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.  Providers may also use third parties to carry out hash matching on their behalf. ICO guidance is clear that where third parties are used, it is for the service provider and that third party to identify their respective roles and obligations under data protection law, and ensure that all the requirements of that law are met.[635] Service providers will also need to comply with the rules on "restricted transfers" in the UK GDPR if transferring personal data outside the UK for the purposes of implementing the measure, ensuring that appropriate safeguards are in place with respect to all such transfers.

4.165 The UK GDPR sets out seven key principles that sit at the heart of data protection law (and which service providers will need to consider as part of ensuring their processing complies with data protection rules). These are:

a) lawfulness, fairness and transparency;
b) purpose limitation;
c) data minimisation;
d) accuracy;
e) storage limitation;
f) integrity and confidentiality (security); and
g) accountability.

4.166 It provides a higher level of protection for the processing of particularly sensitive categories of data such as race, sexual orientation, sex life and health data (known as special category personal data) and criminal conviction and offence data.

4.167 The UK GDPR also 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 Information Commissioner's Office (ICO) has set out that decisions to take down content can (in some cases) have such effects.[636] 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.[637]

4.168 We are satisfied that the processing required by the measure can be carried out in accordance with data protection law. The ICO has published guidance on content moderation and data protection which explains how data protection law applies to content moderation technologies and processes, including where these are solely automated, and provides advice to help service providers comply with the UK GDPR and the Data Protection Act 2018 when utilising these technologies and processes.[638] This includes advising service providers to carry out a data protection impact assessment to assess and mitigate data processing risks.

### The right to privacy under Article 8 ECHR

4.169 We now consider the impact on users' right to privacy under Article 8 ECHR more broadly. An interference with the right to privacy must be in accordance with the law, pursue a

[635] See ICO guidance on controllers and processors. [accessed 29 November 2024].
[636] ICO, 2024, Content moderation and data protection, p.41. [accessed 18 October 2024].
[637] See Article 22(3) of the UK GDPR and section 14 of the Data Protection Act 2018.
[638] ICO, 2024. Content moderation and data protection. [accessed 18 October 2024].

legitimate aim, be proportionate to the legitimate aim and correspond to a pressing social need.

4.170 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. The processing will involve the automated use of an algorithm (the hash function) to create a hash for the purposes of matching to hashes of identified CSAM.

4.171 Review of some detected content by human moderators may have a more significant impact on users' privacy.[639] However, since the measure only applies in relation to content communicated publicly (see paragraph 4.5), the user sharing the content should have a reduced expectation of privacy in connection with that content (compared to, for instance, content shared privately with a small number of other persons).[640]

4.172 Where CSAM is detected, service providers may (in some circumstances) be required (or choose) to report the relevant user to a law enforcement authority (or to a designated reporting body such as NCMEC). For example, US providers are obliged to report to NCMEC under US law when they become aware of child sexual abuse on their services. Relevantly, section 66 of the Act (which is not yet in force) sets out duties for providers of regulated U2U services to report to the NCA detected CSEA content which is not otherwise reported, which may result in information about the user also being provided.[641]

4.173 Aspects of these duties are to be further defined in regulations made by the Secretary of State.[642] However, a report based on a false positive which contained (or resulted in the disclosure of) information about an identified or identifiable individual would be a significant intrusion into that individual's privacy, even if triage processes are in place to ensure that no further action is taken.[643]

4.174 In its consultation response, the ICO said that the accuracy principle in data protection law "requires that [service providers] take all reasonable steps to ensure that the personal data they process is not incorrect or misleading as to any matter of fact". It observed that "the level of accuracy that is appropriate for reports to the NCA (which carries a particular risk of serious damage to the rights, freedoms and interests of a person who is incorrectly reported) and other significant but potentially less harmful actions such as content

[639] Review by human moderators of content accurately detected to be CSAM also represents a significant interference with the privacy rights of the victims it depicts. However, that review forms an important part of ensuring that the overall measure – which aims to protect victims, including from serious impact on their privacy from the further dissemination of CSAM – is proportionate and appropriate for service providers to take for the purposes of complying with their illegal content safety duties. We therefore consider that the intrusion into victims' privacy rights is necessary and that no less intrusive approach would be a suitable alternative.

[640] In its consultation response, the ICO argued that if our guidance on content communicated publicly and privately did not provide "sufficient direction and certainty", there would be "a risk that some services will default to assessing content as being communicated publicly. This would undermine the effectiveness of the privacy safeguard in practice". ICO response to November 2023 Consultation, p.22. We address this in Volume 3, chapter 4: 'Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act'.

[641] In the case of a non-UK provider, the duty is limited to "UK-linked" CSEA content: s.66(2).

[642] Section 67 of the Act requires the Secretary of State to make regulations which will set out the information to be included in reports to the NCA and may also require the retention of user-generated content, user data and associated metadata.

[643] ICO response to November 2023 Consultation, p.13.

takedown" differed, and suggested that the measure should set out that service providers "take into account the importance of minimising false positives being reported to the NCA" when configuring hash-matching technology and when deciding what proportion of detected content human moderators should review.[644] We have adopted this suggestion as a safeguard for users' privacy.

4.175 We have also adopted the ICO's suggestion to specify other code measures as safeguards for users' privacy (including the protection of personal data).[645] We consider that 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.

4.176 Interference with users' privacy could also result from action taken against users by service providers based on a false positive. Such impacts could be significant (for instance, affecting users' ability to participate in economic, social, cultural and leisure activities), and could be especially serious if it were in some way possible for other persons to infer why that action had been taken. 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.

4.177 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.

## Who this measure applies to

4.178 In the November 2023 Consultation we set out our provisional view that the hash-matching measure we had proposed should apply to the following services:

a) large services which are at medium or high risk of image-based CSAM in their risk assessment;
b) other services which are at high risk of image-based CSAM in their risk assessment and have more than 700,000 monthly UK users;
c) services which are at high risk of image-based CSAM which are file-storage and file-sharing services that have more than 70,000 monthly UK users.

4.179 Following the consultation, we have decided to change the scope of the file-storage and file-sharing services that the measure applies to. We have otherwise decided to proceed with the approach we proposed. We conclude that our approach is proportionate considering the scale and severity of CSAM online, our analysis of the effectiveness of the measure, the costs to service providers of implementing it, and its impact on user rights.

We have decided that this measure should apply to:

a) Large services which are at medium or high risk of image-based CSAM;

[644] ICO response to November 2023 Consultation, pp.13-14.

[645] ICO response to November 2023 Consultation, pp.16-17.

b) Services which are at high risk of image-based CSAM and have more than 700,000 monthly active United Kingdom users;[646]

c) Services which are at high risk of image-based CSAM and are file-storage and file-sharing services.[647]

4.180 We have decided to apply the measure to all file-storage and file-sharing services which are at high risk of image-based CSAM (and are therefore removing the user threshold proposed in the November 2023 Consultation), for the reasons explained below.

### How service providers should assess their risk level for image-based CSAM

4.181 As part of this Statement, we have published Risk Assessment Guidance for Service Providers. Part 3 of that guidance includes guidance to help providers of U2U services to make an assessment of their level of risk for image-based CSAM. As set out in Table 13.1 of the guidance, we would normally expect a service to be assessed as **high risk** where:

- there is evidence that image-based CSAM has been present to a significant extent on the service; or
- the service enables images or videos to be generated, uploaded or shared and is:
  - a file-storage and file-sharing service, or
  - an online adult (pornography) service, or
  - a lot of the specific risk factors specified in the guidance have been identified.[648]

4.182 We would normally expect a service to be assessed as **medium risk** where:

- there is evidence that image-based CSAM has been present on the service, but not to a significant extent; or
- the service enables images or videos to be generated, uploaded or shared and several of the specific risk factors specified in the guidance have been identified.[649]

4.183 Service providers should refer to Table 13.1 of the guidance for more information, including as to when it could be appropriate to assess at a lower risk level.

4.184 Overall, and as mentioned in paragraph 4.19, stakeholders expressed general support for our consultation proposals for which services the measure should apply to. However, some stakeholders did express views on applying the measure to more U2U services.[650]

[646] As calculated in accordance with the methodology set out in the Codes of Practice. See the 'Our approach to developing Codes measures' chapter for more information.

[647] For the purposes of this measure a file-storage and file-sharing service is defined as: a service whose primary functionalities involve enabling users to (i) store digital content, including images and videos, on the cloud or dedicated server(s); and (ii) share access to that content through the provision of links (such as unique URLs or hyperlinks) that lead directly to the content for the purpose of enabling other users to encounter or interact with the content. This definition is consistent with that used in the Risk Profiles and the Register of Risks, and Service Risk Assessment Guidance.

[648] The specific risk factors specified in Table 13.1 of the guidance are: (a) child users, (b) social media services; (c) messaging services; (d) discussion forums and chat rooms; (e) group messaging; (f) livestreaming; (g) direct messaging; (h) encrypted messaging; (i) users can share images or videos without creating a user account.

[649] See footnote above.

[650] Generally, respondents suggested widening the scope of services across all the Codes measures.

### File-storage and file-sharing services

4.185 This subsection explains why we have decided to change the scope of the file-storage and file-sharing services that the measure applies to from the approach proposed in the November 2023 Consultation.

4.186 In the November 2023 Consultation, we proposed that this measure apply to file-storage and file-sharing services which are at high risk of image-based CSAM and have more than 70,000 monthly UK users.[651]

4.187 As explained in that consultation, we proposed a lower threshold for file-storage and file-sharing services because we considered that they play a particularly central role in the dissemination of known CSAM.[652] We also noted that only a handful of file-storage and file-sharing services reached more than 700,000 monthly UK users. Therefore, if we set the same user numbers threshold for file-storage and file-sharing services as for other service types, this would mean that the proposed hash-matching measure would not apply to many file-storage and file-sharing services, including many file-storage and file-sharing services that have a significant amount of CSAM on them.

4.188 At that time, we also outlined various other options we considered before reaching our provisional view.[653] This included an option to recommend the measure to all services with a medium or high risk of image-based CSAM, regardless of user numbers. One reason why this was not our preferred option was because there are only a finite number of providers of hash matching databases, and we understood these services only have a finite amount of capacity to provide their services to clients. Therefore, we were concerned that if we applied the measure too widely, the database ecosystem would not be able to cope with levels of demand, at least in the short term.

4.189 We received several responses from stakeholders regarding the scope of the measure and the range of file-sharing and file-storage service providers that would be required to implement the measure. We detail these responses in paragraphs 4.48 to 4.50. In response to this feedback, we reviewed our evidence base relating to file-storage and file-sharing services, with a view to whether we should alter the scope of file-sharing and file-storage services that the measure applies to. As well as reviewing responses to our consultation, we have spoken with experts from law enforcement and other fields, and we have conducted desk research into the market for file-storage and file-sharing services. The evidence confirms that many file-storage and file-sharing services are at exceptionally high risk of image-based CSAM, including services with fewer than 70,000 monthly UK users. This suggests that we would not be able to achieve our policy objectives in relation to combatting CSAM if we kept the 70,000 user threshold in place.

- File-storage and file-sharing services are vulnerable to be exploited by perpetrators because their design focus and functionalities enable the storing and sharing of image-based CSAM. The risk of harm from illegal content on these services does not necessarily scale with the number of users. Many file-storage and file-sharing services allow users to store and share vast numbers of files, meaning that small numbers of

[651] This was much lower than the 700,000 user threshold we proposed should apply to other services with a high risk of image-based CSAM. We also proposed that the measure apply to all large services at medium or high risk of image-based CSAM.

[652] November 2023 Consultation, Volume 4, paragraphs 14.112 to 14.114.

[653] November 2023 Consultation, Volume 4, paragraph 14.105.

users could use the service to disseminate CSAM at scale. Moreover, as discussed in the Register, some services with smaller user bases offer users specific functionalities which may not be available on services with larger user bases, such as the ability to post content without a registered account. As such, perpetrators may target these kinds of smaller services to exploit such functionalities.

- The evidence supports that there are many file-storage and file-sharing services which have a significant amount of image-based CSAM. For example, in 2023, INHOPE reported that approximately 39% of the CSAM it detected was hosted by 'image hosts'. It also reported that 5% of the CSAM detected was hosted by 'file hosts', although this figure has been as high as 26% in previous years. INHOPE attributed this reduction from previous years to difficulties in detecting illegal content on file hosting services that require payment rather than due to a reduction in risk.[654] The IWF found that 89% of images or videos detected of livestreamed child sexual abuse were stored on an 'image-hosting service'. Other types of file-storage and file-sharing services such as 'cyberlockers' and 'image stores' made up 4% and 1% of cases respectively of the child sexual abuse imagery that the IWF reviewed in 2023.[655] This evidence demonstrates the significant exploitation of file-sharing and file-storage service providers for the purpose of disseminating CSAM.
- The evidence supports that many file-storage and file-sharing services with a significant amount of image-based CSAM have fewer than 70,000 monthly UK users. In 2021, C3P published a report which listed tens of services on which it had detected over 5,000 images of pre-pubescent CSAM, post-pubescent CSAM, and harmful abuse.[656] The report concluded that "the vast majority of [services] that have received removal notices…have been image hosting providers or file hosting services", which would be included in our definition of file-storage and file-sharing services. Moreover, many of these file-storage and file-sharing services have fewer than 70,000 monthly UK users. This is consistent with the views of internal experts who understand how the file storage and sharing market operates and of experts from law enforcement and NGOs who understand how CSAM manifests online.[657]

4.190 Based on this evidence, we considered whether the measure should also be applied to file-storage and file-sharing services with fewer than 70,000 monthly UK users.

4.191 We considered the costs and benefits of applying the measure to smaller file-storage and file-sharing services.

- We recognise that small file-storage and file-sharing services are likely to have constrained resources which could affect their ability to implement hash-matching. However, small file-storage and file-sharing services can access hash-matching technology at relatively low cost through API-based solutions. For example, C3P described that as part of Project Arachnid, it provides a no-cost approach for operators, which is to register for access to its API. Under this approach, an operator simply submits incoming user-uploaded media to the API and receives a response as to

[654] INHOPE are a global network of organisations working to tackle CSAM. Source: INHOPE, 2023. Annual report 2023, p.39. [accessed 7 August 2024].
[655] IWF, 2023. Annual report 2023. [accessed 7 August 2024].
[656] C3P, 2021. Project Arachnid: Online availability of child sexual abuse material. [accessed 08 November 2024].
[657] [✂].

whether the image is known CSAM or not.[658] We have updated our estimate for the minimum annual tech cost likely to be incurred by services, since the November 2023 Consultation. As mentioned above in (see paragraph 4.111), we now assume this cost to start from £1,000, based on IWF's membership fees, which includes access to its hash list and Microsoft's PhotoDNA.[659] We are also aware that tech costs may be lower where a service is deemed to have less ability to pay by a relevant safety tech provider, and that there are other low cost or free options available in the market.

- As set out above, we recognise that services will incur other costs in addition to tech, including build costs, maintenance costs, and moderation costs. These costs are quantified in Annex 5 and stacked up against one benefit of hash matching: reducing contact abuse. We find that even if we only quantify the benefit of hash matching from reducing contact CSA, this benefit is likely to far exceed the direct costs for a file-storage and file-sharing service that reaches 70,000 monthly UK users. This suggests that the benefits of doing hash matching would far exceed the costs for many file-storage and file-sharing services with fewer than 70,000 monthly UK users. This is particularly true for file-storage and file-sharing services with a high risk of image-based CSAM. These services with a high risk of CSAM are likely to incur relatively high moderation costs, but this is because they are likely to find CSAM on their service and so the moderation costs would be outweighed by the direct and indirect benefits of removing known CSAM from the service. See Annex 5 for further details.

4.192 We also considered other implications of applying the measure to smaller file-storage and file-sharing services. These include impact on the wider hash-matching ecosystem, the risk of perpetrators switching to smaller services and fluctuation of user numbers on smaller file-storage and file-sharing services.

- In the November 2023 Consultation, we explained our concern that applying the measure to all services at high or medium risk of image-based CSAM, regardless of their user base size, could put significant pressure on the wider hash-matching ecosystem and the ability of relevant providers of hash databases to meet the significant increase in demand, at least in the short term.[660] In responses to the consultation, several hash database providers indicated they have capacity to support a much greater number of service providers, including file-sharing and file-storage services, in implementing the measure. C3P said that its Project Arachnid tool already "handles up to 10s of millions of images every day from smaller providers".[661] The IWF said: "We want to see the hashing and URL provisions embedded as widely as possible across industry for the maximum impact to be achieved. Recognising the increased costs to micro, small and medium sized businesses, we could support recommendations that they are given longer to prepare, maybe a period of 12-18 months but if they are medium to high risk of harm, they should be required and in scope of mitigation measures".[662] Overall, we are satisfied that the capacity constraint is less of a concern than we thought when consulting on the measure.

[658] C3P response to November 2023 Consultation, p.19; [✂].

[659] IWF, 2024. Membership fees. [accessed 4 November 2024]; and IWF, 2024. Image hash list. [accessed 4 November 2024].

[660] See November 2023 Consultation, Volume 4, paragraph 14.110.

[661] C3P response to November 2023 Consultation, p.19.

[662] IWF response to November 2023 Consultation, p.21.

- Another factor we considered was the risk of perpetrators switching away from file-storage and file-sharing service that are in scope of the measure to similar services that are also high risk for image-based CSAM but that do not do hash matching. File-storage and file-sharing services are relatively substitutable in terms of the basic functionalities they offer, and users can switch between such services with relative ease. This risk is especially pronounced among smaller file-storage and file-sharing services, whose users are not attracted to the service because of the network effects resulting from their service being used by a large proportion of the UK population. This means that harm relating to CSAM on a small file-storage and file-sharing service could increase sharply in a short space of time. We note that this risk – of perpetrators switching to smaller services - would remain regardless of where a user threshold was set, and so would only be eliminated if the user threshold on file-storage and file-sharing services was removed.
- A further issue with setting a low user threshold is that user numbers can fluctuate substantially for smaller file-storage and file-sharing services. For example, our analysis of UK audience data found that, over a period of four months (between May and August 2024), it was not uncommon for the estimated number of users for file-storage and file-sharing services to fluctuate by 30% to 80%.[663] These types of fluctuations generally do not reflect changes to the risk of a service for image-based CSAM, and so services could fall in and out of scope of our measure despite continuing to be of high risk for CSAM.

4.193 In summary, our analysis suggests that: (i) if we left the proposed 70,000 user threshold for filesharing services in place our ability to combat CSAM on these services would be significantly reduced; (ii) our initial concerns about capacity in the database ecosystem were somewhat overstated; and (iii) given the severity of the harm CSAM causes and the prolific amounts of CSAM that even very small filesharing services can host, it would be proportionate to apply the measure to all file-storage and file-sharing services at high risk of image-based CSAM, including those with fewer than 70,000 monthly UK users.

4.194 As explained above, we remain of the view that the evidence supports applying a different approach to file-storage and file-sharing services than to some other service types (while also recognising that other service types can be at high risk of being used to commit or facilitate offences relating to CSAM). However, our decision only removes the user threshold for file-storage and file-sharing services that are at high risk of image-based CSAM. We have also revised the guidance for assessing risk for image-based CSAM in our Risk Assessment Guidance to make clear that the risk level table does not set out definitive criteria and that strong reasons could justify assessing risk at a lower level. We consider that this appropriately addresses the criticisms made by some service providers, as summarised at paragraphs 4.48 to 4.50.

### Applying the measure to more U2U services

4.195 We received feedback from two stakeholders (as outlined in paragraph 4.44) that we should clarify our approach to reviewing the capacity of hash database providers, so that we can broaden the scope of our measure over time to include for more services. We will continue to look at the scope of the measure, including the relevant factors, as part of our

[663] This analysis was based on user numbers provided by Similarweb, which is widely used by industry to estimate website traffic.

work on the Codes. We also reviewed feedback about whether the measure should apply to other types of services, as outlined in paragraphs 4.43 and 4.45.

- Some stakeholders suggested the scope of the measure should be broadened to include more services or different types of services, and that user thresholds should be regularly reviewed or removed entirely.
- Some service providers contended that our proposed measure would apply too broadly and that the recommended approach to risk assessment should be revised.
- Several stakeholders argued in favour of expanding the scope of the measure to include providers of smaller services, whilst other stakeholders suggested the costs associated with hash-matching would make the measure difficult for providers of smaller services to implement.

4.196 For reasons set out in this chapter and Annex 5, our view at this time is that it is proportionate to recommend hash matching to services with a high risk of image-based CSAM that have over 700,000 monthly UK users and services with a medium risk of image-based CSAM that have over 7 million monthly UK users. We reference some types of services in the Risk Assessment Guidance as being especially likely to be high or medium risk for image-based CSAM.

4.197 Given the severity of CSAM, we consider that it would also likely be proportionate to recommend that high risk services with fewer than 700,000 monthly UK users hash match even where they are not file-storage and file-sharing services. However, although we now believe that the hash-matching ecosystem will have enough capacity to accommodate all high risk file-storage and file-sharing services, we remain concerned that if we introduced an expectation that all high risk services use hash matching, the ecosystem could be overwhelmed with demand. At this point, we have therefore decided to include all high risk file-storage and file-sharing services in the scope of this measure irrespective of size, but leave the other thresholds we proposed in the November 2023 Consultation in place.

4.198 Other relevant measures may apply to high or medium risk services which do not meet the relevant user thresholds, including the measures relating to content moderation, governance processes and tracking evidence of new and increasing illegal harm.[664] All service providers must have a content moderation function that allows for the swift takedown of illegal content of which they are aware. If a service provider tracks and observes an increase in CSAM being taken down from its service, that may increase the risk level and mean that the service comes into scope of this measure.

4.199 Overall, we have revised the scope of this measure to reflect the risk posed by file-sharing and file-storage providers by applying to all high-risk services. We have not changed the other thresholds for other U2U service providers which remain consistent with our approach proposed in the November 2023 Consultation.

4.200 Therefore, we have concluded this measure is proportionate for the services set out at the start of this section.

[664] See governance measures ICU A2, ICU A3, ICU A5, ICU A6, and ICU A7 and content moderation measures ICU C1-C8. (Some of these measures only apply to large services and/or multi-risk services.)

## Conclusion

4.201 We consider this measure to be an effective means to help prevent users of U2U services from encountering CSAM on the service, with substantial benefits. Having considered the costs, risks and associated impacts on the rights of users, we consider it to be a proportionate safety measure to recommend providers (of the services to which we have decided to apply the measure) take. The measure is recommended for the purpose of complying with (in particular) the duties under section 10(2)(a) to use proportionate measures to prevent individuals from encountering priority illegal content by means of the service, and under section 10(3)(a) to use proportionate systems and processes designed to minimise the length of time for which any priority illegal content is present.

4.202 We recognise that the costs to service providers of this measure will in many cases be significant. However, in light of the expected effectiveness of the measure and the benefits associated with removing CSAM, which causes particularly serious harm, we consider it appropriate for providers to incur those costs.

4.203 We recognise the potential adverse impacts on users' rights from the use of such proactive technology. However, we have mitigated these by designing the measure to incorporate a number of safeguards for the protection of users' rights to freedom of expression and privacy. We further consider that the adverse impacts arising from the measure are proportionate to its aims and that there is no less intrusive way of achieving those aims.

4.204 We have therefore decided to include the measure in the CSEA Code of Practice. It is referred to as ICU C9.

# Measure on detecting and removing content matching listed CSAM URLs

4.205 In the November 2023 Consultation, we proposed that certain U2U service providers use URL detection technology to detect and remove direct matches in user-generated content communicated publicly on the service to a list of known CSAM URLs, for the purpose of complying with their illegal content safety duties under section 10(2) and (3) of the Act. We proposed that service providers analyse content already present on the service (within a reasonable time), as well as new content uploaded to the service or which a user seeks to upload (before or as soon as practicable after it can be encountered by other users). We also proposed that any content detected to be a CSAM URL be swiftly taken down (or prevented from being generated, uploaded, or shared). We proposed that this measure should apply to:

- large services which are at medium or high risk of CSAM URLs in their risk assessment; and
- other services which are at high risk of CSAM URLs in their risk assessment and have more than 700,000 monthly United Kingdom users.

4.206 We proposed that these service providers should source an appropriate list of known CSAM URLs from a third party that (1) has expertise in the identification of CSAM, and (2) meets other criteria specified in the measure. We said that the service should compare content to the latest version of the list.

4.207 We explained that the aim of this measure was to reduce users' exposure to CSAM on other services. The online circulation of CSAM causes significant and potentially lifelong harm, including re-traumatising victims and survivors of sexual abuse.[665]

## Summary of stakeholder responses

4.208 A range of stakeholders, including providers of regulated services, governments, law enforcement, academics, and civil society organisations, expressed broad support for this proposed measure.[666] Similarly to the feedback on our proposed hash-matching measure, these stakeholders expressed general support for the use of automated systems to remove priority illegal content due to the volume of content posted and shared online, or specifically approved of using such systems to remove CSAM given the seriousness of the harm it causes. Some of these stakeholders also indicated that human moderation alone is not a feasible option for moderating many services due to the volume of content.

4.209 We received several responses expressing support for the proposed measure, including more specific elements of the measure. For example:

- IICSA Changemakers agreed that URL detection tools are critical for ensuring CSAM can be proactively detected, removed, and reported.[667]
- The NSPCC strongly supported our proposals on URL detection, citing the negligible risk to user privacy for non-victims.[668]
- The Welsh Government expressed strong support for the measure, citing the importance of protecting children and young people from CSAM.[669]
- Segregated Payments Limited were supportive of the proposed measure, which it stated was "in line with industry best practice" and readily available technologies.[670]

4.210 Several stakeholders, however, commented on or expressed concerns about specific elements of the proposed approach and/or suggested changes to improve the measure. This feedback related to:

- the effectiveness of the proposed measure;
- associated costs and risks with the proposed measure;
- implications for users' rights; and
- who we proposed the measure should apply to.

[665] See Register of Risks chapter titled 'CSEA'.

[666] Are, C. response to November 2023 Consultation, p.8; Barnardo's response to November 2023 Consultation, p.16; BILETA response to November 2023 Consultation, p.11; C3P response to November 2023 Consultation, p.17; CELE response to November 2023 Consultation, p.9; Children's Commissioner response to November 2023 Consultation, p.22; IICSA Changemakers response to November 2023 Consultation, p.4; ICO response to November 2023 Consultation, pp.2; IJM response to November 2023 Consultation, p.11; IWF response to November 2023 Consultation, p.31; [✂]; Microsoft response to November 2023 Consultation, p.11; [✂]; Nexus response to November 2023 Consultation, p.10; NSPCC response to November 2023 Consultation, p.23; OnlyFans response to November 2023 Consultation, p.5; Pinterest response to November 2023 Illegal Harms Consultation, p.8; Segregated Payments Ltd response to November 2023 Consultation, p.8; SPRITE+ (York St John University) response to November 2023 Consultation, p.10; Welsh Government response to November 2023 Consultation, p.3; WeProtect Global Alliance response to November 2023 Consultation, p.13; X response to November 2023 Consultation, p.4.

[667] IICSA Changemakers response to November 2023 Consultation, p.4.

[668] NSPCC response to November 2023 Consultation, pp.23-24.

[669] Welsh Government response to November 2023 Consultation, p.3.

[670] Segregated Payments Limited response to November 2023 Consultation, p.8.

4.211 We outline this feedback in the following paragraphs.

## Feedback on the effectiveness

4.212 Several of the points raised by stakeholders related to the effectiveness of the proposed approach and its potential benefits for reducing the spread of CSAM. The IWF described to the measure as "extremely effective at dealing with the issue of known (previously detected) [CSAM]". It referred to data provided to it in 2020 showing that three companies which implement the IWF's URL list across their UK networks had blocked and filtered 8.8 million attempts originating from the UK to access webpages on its blocking list in one month. It also referred to an announcement by one of its members, Converge (a fibre broadband and technology provider) that it had blocked 9.8 billion requests to be connected to sites (not limited to CSAM) in 2023, compared to around 1.9 billion requests in 2022. Converge attributed this rise (in part) to adding 198,000 URLs and domains associated with illegal activities to its blocking list, including through its partnership with the IWF.[671]

4.213 WeProtect noted that CSAM is often published and hosted in different jurisdictions, which can complicate the evidence-gathering process. It supported recommending that service providers use URL lists to facilitate the detection of CSAM, stating that the IWF URL list "is cited as a helpful tool in identifying potential harms and blocking access to illicit webpages and material." It further indicated that the IWF is constantly updating and reviewing its list (twice a day) of URLs to support detection and removal of CSAM. WeProtect also indicated that Project Arachnid – a project run by Canada's reporting service CyberTipline as part of C3P – has been highlighted as an effective technology to combat link-sharing. It said that the project identifies child sexual abuse material by crawling specific publicly accessible URLs reported to CyberTipline, as well as URLs on the surface web and dark web that are proven or known to host CSAM.[672] Comparatively, the Association of British Insurers (ABI) expressed concerns that URL detection does not stop users from creating new URLs with similar web addresses (this point was raised in the context of online scams, however, we consider it to still be relevant).[673]

4.214 Some stakeholders suggested that the proposed approach needs to provide flexibility for service providers by not prescribing specific technology to detect URLs, claiming that this would improve the effectiveness of the measure. Meta recommended that the measure should not impose the use of specific technology but should allow service providers to choose from a variety of solutions. It also noted that URL detection can be done in stages, from simple text matching, all the way to crawling the content of the link and hashing media, and that service providers can refer to industry initiatives to access a shared corpus of CSAM URLs to feed their own URL detection technology.[674] Similarly, Stop Scams UK advocated for a "dynamic and flexible approach" to prescribing specific systems and techniques that service providers should use for content moderation.[675] We address this input in the 'Benefit and effectiveness' section.

[671] IWF response to November 2023 Consultation, p.9.

[672] WeProtect Global Alliance response to November 2023 Consultation, pp.15-16.

[673] ABI responses to November 2023 Consultation, p.3.

[674] Meta response to November 2023 Consultation, annex, p.10.

[675] Stop Scams UK response to November 2023 Consultation, p.11.

4.215 We also received mixed stakeholder feedback on the use of URL detection, including fuzzy matching, and its effectiveness for identifying and removing URLs.[676] Ukie suggested that the main costs for URL detection are associated with engineering and labour, and that fuzzy matching has limited value due to the low prevalence of CSAM URLs on its members' networks.[677] We address this input in the ‘Benefits and effectiveness’ and ‘Costs and risks’ sections.

4.216 Some stakeholders recommended the use of ‘splash pages’ as an approach to removing content from online services and explained that splash pages can notify users that the content they are attempting to view is illegal and can provide information on where they can seek confidential help or speak to a professional. The Lucy Faithfull Foundation recommended the use of a splash page to signpost users attempting to view CSAM. It further argued that splash pages help to prevent offending by “informing people who try to access blocked URLs about the illegality of viewing CSAM, the harm sexual abuse causes to children and the consequences of offending for themselves and their families. In addition, the splash pages direct the user to confidential and anonymous help to change their behaviour from Stop It Now services”.[678] We also received a response from Barnardo’s that expressed concerns that our approach does not reflect best practice and indicated the “IWF recommend that a splash page is served”.[679] We address this input in the ‘How this measure works’ and ’Benefits and effectiveness’ sections.

4.217 We received responses from third-party URL list providers indicating their ability to support service providers in implementing the proposed measure. The IWF noted its willingness and preparedness to provide access to its CSAM URL list (and hash database) in accordance with the proposed measure. It also suggested we consider accrediting technological solutions based on high-quality data provided by organisations such as the IWF.[680] Similarly, C3P indicated that its tool ‘Shield by Project Arachnid’ could facilitate URL matches at no cost for service providers.[681] [682] We also received input from service providers that indicated these lists were easily accessible and effective. [✂].[683] Pinterest recommended that Ofcom provide service providers with lists of known CSAM domains to support detection and removal of CSAM.[684] We address this input in the ‘Benefits and effectiveness’ section.

## Feedback on the associated costs and risks

4.218 Glitch raised the risk of security vulnerabilities of URL lists and the need to strengthen safeguards within the measure to protect against exploitation and/or unauthorised access to them. It stated that the measure does not address how security vulnerabilities of URL lists (and hash databases) may disproportionately impact women and girls. It further questioned how the measure will account for any loopholes that enable the exploitation of

[676] Fuzzy matching will surface matches that are similar (for example, those which end ‘dotcom’, instead of ‘.com’).

[677] Ukie response to November 2023 Consultation, p.21.

[678] Lucy Faithfull Foundation response to November 2023 Illegal Harms Consultation, p.5.

[679] Barnardo’s response to November 2023 Consultation, p.17.

[680] IWF response to November 2023 Consultation, p.4.

[681] C3P response to November 2023 Consultation, p.19.

[682] This should be caveated with the fact the legislative definitions of CSAM in Canada differ to those in the UK, and we will need investigate if their database can conform to our principles.

[683] [✂].

[684] Pinterest response to November 2023 Consultation, p.8.

security and privacy protections.[685] Other stakeholders also suggested strengthening the security provisions by recommending service providers develop mechanisms to mitigate security risks, including security protocols for hash functions or databases to protect against malicious activity.[686] We consider this input is also relevant to URL lists. This concern is addressed in the 'Costs and risks' section.

4.219 We also received input from OSTIA, although in the context of the hash-matching measure, noting the risk of biases in hash databases. It suggested we explicitly recommend that third-party database providers should avoid "systematic bias within their control", arguing that databases should "determine addition of content solely based on whether or not it is CSAM and ensure minimisation of bias in processes making that determination".[687] Glitch also shared concerns that automated moderation technologies "may perpetuate biases present in training data or design, leading to disproportionate moderation outcomes for women and girls".[688] This concern is addressed in the 'Costs and risks' section.

4.220 Some stakeholders raised concerns regarding the risk that the proposed measure could lead to over-moderation which, in turn, could result in an increase in human moderation. Microsoft expressed the view that the current form of URL detection technology is insufficient to enable implementation of the proposed measure without the risk of over-moderation. It argued that the measure could result in a significant number of moderation decisions that would be later overturned on appeal, noting that this would likely require a large increase in the need for human moderation to investigate the open internet for further context. It also commented that the length of time required to investigate each case could result in slower resolution of CSAM cases generally, increasing the wellness risks to moderators.[689] [✂].[690] This concern is addressed in the 'Costs and risks' section.

4.221 Some stakeholders expressed concerns that the costs associated with the proposed measure would impact the wider ecosystem. Yoti questioned why there were no cost estimates for preventative measures that address CSAM and suggested we consider the costs to law enforcement and civil society generated by this measure. It further argued that the cost implications would continue to spiral "if preventative measures are not deployed across the ecosystem, with regulators working in conjunction with payment processors and ad networks as well as platforms".[691] We consider this concern in the 'Costs and risks' section.

4.222 Similarly, Protection Group International argued that URL detection is time-consuming for service providers, law enforcement agencies, and other relevant organisations.[692] We consider this comment in the 'Benefits and effectiveness' section.

[685] Glitch response to November 2023 Consultation, p.9.

[686] Global Partners Digital response to November 2023 Consultation, pp.14-15; INVIVIA response to November 2023 Consultation, p.15; Meta response to November 2023 Consultation, confidential annex, p.10; Microsoft (confidential) response to November 2023 Consultation, pp.13-14; Proton response to November 2023 Consultation, p.6.

[687] OSTIA response to November 2023 Consultation, p.14.

[688] Glitch response to November 2023 Consultation, p.8.

[689] Microsoft response to November 2023 Consultation, p.11.

[690] [✂].

[691] Yoti response to November 2023 Consultation, p.11.

[692] Protection Group International response to November 2023 Consultation, p.7.

## Feedback on the implications for users' rights

4.223 Some stakeholders expressed concerns regarding the impact of the proposed measure on users' rights. Big Brother Watch stated that automated moderation systems often result in over-removal of content which, in turn, would risk infringing on users' rights to freedom of expression.[693] The ICO also highlighted the risks that automated processing could pose to the rights of users, explaining that the moderation of content using automated means still has data protection implications for users whose content is being scanned. It expressed concerns with the privacy impact assessment set out in the November 2023 Consultation (in which we set out our provisional view that the privacy risk arising from automated scanning was minimal) and was of the view that the measure needed to be supported by a fuller impact assessment which takes account of data protection impacts. The ICO also suggested that privacy safeguards in the proposed automated content moderation measures should be expanded to cover data protection requirements, including transparency, purpose limitation, data minimisation, and accuracy, as well as compatibility with the requirements in Article 22 of the UK GDPR (which places restrictions on solely automated decision-making based on personal information). The ICO suggested service providers should be required to take into account the importance of minimising incorrect reports of CSAM to the NCA when configuring technical accuracy and deciding on the proportion of material appropriate for human review.[694] This concern is addressed in the 'Rights impact' section.

## Feedback on who this measure applies to

4.224 Some stakeholders who responded to the November 2023 Consultation expressed support for the scope of service providers covered by the proposed measure.[695] However, others suggested that the scope of the measure should be expanded to include more services, including all high-risk services, smaller services, and any services with at least a medium risk of CSAM.[696] We also received feedback from one individual indicating that the measure's scope should be narrower and should not apply to smaller services. [697]

### Broadening scope to smaller services

4.225 C3P suggested that a range of smaller service providers are being used to share CSAM, which are not in scope of this measure. In particular, it suggested that services tailored to children, which are used by a smaller portion of the population and may not be in scope, can be exploited by adults seeking to cause harm to the users. C3P suggested that such services are less likely to meet a user threshold that is based on the entire population and propose that a more appropriate alternative would be the number of child users.[698] Barnardo's argued that economic-based proportionality concerns should not prevent the

[693] Big Brother Watch response to November 2023 Illegal Harms Consultation, p.8.

[694] ICO response to November 2023 Consultation, pp.2, 10-17.

[695] [✂]; Marie Collins Foundation response to November 2023 Consultation, p.8; Welsh Government response to November 2023 Consultation, p.3; WeProtect Global Alliance response to November 2023 Consultation, p.27.

[696] Barnardo's response to November 2023 Consultation, pp.13, 16; C3P response to November 2023 Consultation, pp.4, 12; Children's Commissioner response to November 2023 Consultation, p.19; IWF response to November 2023 Consultation, p.31; Marie Collins Foundation response to November 2023 Consultation, p.11; [✂]; NSPCC response to November 2023 Consultation, p.24; The Cyber Helpline response to November 2023 Consultation, p.10.

[697] Name withheld 3 response to November 2023 Illegal Harms Consultation, p.12.

[698] C3P response to November 2023 Consultation, pp.4, 12.

measure from effectively being implemented across all services. It disagreed that smaller service providers equate to less harm, suggesting that the proposed measure as a result did not capture many high-risk services, such as gaming services.[699] The NSPCC also suggested that the measure should be applied to "smaller services (700,000 users) with a medium risk of CSAM (rather than just a high risk)".[700] In addition, [✂] argued that if low-risk platforms for child sexual abuse could still implement safety standards, technology, and processes (that are low cost), they should be encouraged to do so.[701]

### Broadening scope to all services

4.226 The Cyber Helpline suggested that "content moderation codes relating to CSAM should apply to all services in scope".[702] The Marie Collins Foundation stated that user number thresholds should be constantly reviewed as services that are high or medium risk for image-based CSAM (regardless of size) should implement [ACM].[703] The IWF proposed that the measure "should apply to all services at medium to high risk of one type of harm", implying that there should not be any user reach threshold for inclusion.[704] The Children's Commissioner recommended that "child safety measures should be applied to all user-to-user services that children may use in order to avoid loopholes that are exploited by unregulated services".[705] [✂].[706]

## Our decision

4.227 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation. We have made a number of relatively minor amendments to the measure in response to the feedback set out in the 'Summary of stakeholder feedback' section. These include:

- clarifying that service providers may use more than one URL list;
- addressing potential risks relating to bias, by providing that service providers should ensure that the arrangements in place for identifying and assessing suspected CSAM URLs for potential inclusion on the list do not plainly discriminate on the basis of protected characteristics (such as sex or race);
- specifying that service providers should ensure an appropriate policy is in place, and measures are taken in accordance with that policy, to secure URL lists from unauthorised access, interference or exploitation; and
- specifying other Code measures which act as safeguards for users' right to freedom of expression and privacy, including allowing users to appeal against the takedown of content.

4.228 Our measure therefore sets out that:

- certain U2U service providers should detect and remove content communicated publicly on the service which matches a URL on a list of URLs previously identified as

[699] Barnardo's response to November 2023 Consultation, p.13.
[700] NSPCC response to November 2023 Consultation, p.24.
[701] [✂].
[702] The Cyber Helpline response to November 2023 Consultation, p.10.
[703] Marie Collins Foundation response to November 2023 Consultation, p.11.
[704] IWF response to November 2023 Consultation, p.31.
[705] Children's Commissioner response to November 2023 Consultation, p.19.
[706] [✂].

hosting CSAM. The measure applies to a provider in respect of each service it provides that:

- has more than 700,000 monthly active United Kingdom users[707] and is at high risk of CSAM URLs; or
- is a large service and is at medium or high risk of CSAM URLs.

4.229 The measure can be found in full in our Illegal Content Codes of Practice for U2U services, within which we refer to this measure as ICU C10. It forms part of the CSEA Code of Practice.

## Our reasoning

### How this measure works

4.230 This measure sets out that providers of certain services should use technology to detect content that matches listed “CSAM URLs” and take down that content. For these purposes, the measure sets out that, for the technology to be effective, it should:

a) compare analysed content to one or more lists of CSAM URLs sourced from a person (or persons) [708] with expertise in identifying CSAM and who meet requirements designed to ensure the list of URLs is accurate and effectively maintained (as explained further below); and

b) detect content as a match for a listed CSAM URL where (1) it is a direct match for a listed URL or (2) it is a URL that contains a listed domain. The measure also specifies that for these purposes it does not matter whether the content includes an access protocol (such as https://).

4.231 A ‘CSAM URL’ is defined as a URL at which CSAM is present,[709] or a domain which is entirely or predominantly dedicated to CSAM.[710] In most cases, we would expect lists to be at URL (not domain) level, so that illegal content can be specifically targeted. However, we recognise that, where a list includes a domain that is predominantly dedicated to CSAM, URLs at that domain which do not themselves include CSAM would be affected. We consider this to be proportionate given the clear risk that users accessing those URLs will go on to encounter CSEA content on other pages at that domain.

4.232 The content in scope of this measure (“relevant content”) is any regulated user-generated content in the form of written material or messages (including hyperlinks) that may be encountered by United Kingdom users of the service and is communicated publicly by means of the service.[711] The analysis can also take place at the point of upload (i.e. before the material is communicated to other users of the service).

[707] As calculated in accordance with the methodology set out in the Codes of Practice. See the ‘Our approach to developing Codes measures’ for more information.

[708] We refer to such persons as an organisation for ease of reference.

[709] As explained in our Illegal Content Judgements Guidance, a link to CSAM should usually itself be considered CSAM, and therefore a URL containing link(s) to CSAM would usually be a CSAM URL for the purposes of this measure.

[710] The measure provides that a domain is “entirely or predominantly dedicated to CSAM” if the content present at the domain, taken overall, entirely or predominantly contains CSAM (such as indecent images of children) or content related to CSEA content.

[711] See Ofcom’s Guidance on content communicated ‘publicly’ and ‘privately’ under the Online Safety Act.

### Selection of a URL list

4.233 As mentioned above, the measure sets out that service providers should source one or more lists of CSAM URLs from an organisation with expertise in identifying CSAM. We have revised the measure to make clear that service providers may use more than one list, reflecting that (for instance) the IWF maintains a separate URL list for "non-photographic imagery" which is assessed as illegal.[712] This makes clear that service providers may use multiple lists in combination, which can enable more links to illegal content to be removed and so provide users with more effective protection.

4.234 The measure sets out requirements that should be met for a list to be appropriate to use for the measure. These include:

- the organisation from which the list is sourced has arrangements in place to identify suspected CSAM URLs, and secure (so far as possible) that they are correctly identified as CSAM URLs before being added to the list.
- an additional requirement that these arrangements for identifying or assessing suspected CSAM URLs do not plainly discriminate on the basis of protected characteristics (within the meaning of Part 2 of the Equality Act 2010), such as sex or race. This change has been made in response to comments from stakeholders about the potential risk of bias in databases or lists of CSAM (for example, if CSAM relating to boys or girls were to be systematically excluded).

4.235 The measure also includes requirements to ensure that the list is effectively maintained, and its integrity assured. The organisation is required to have arrangements are in place to:

- Regularly update the list with identified CSAM URLs.
- Regularly review listed CSAM URLs and remove any which are no longer CSAM URLs (i.e. where the CSAM at the URL has been taken down).
- Secure the list from unauthorised access, interference or exploitation.

### Detection and removal of URLs

4.236 The measure provides that, where technically feasible, service providers should ensure technology is used effectively to analyse relevant content to assess whether it consists of, or includes, content matching a listed CSAM URL. As explained earlier in this chapter, the measure will only apply where it is technically feasible for a provider to implement it.

4.237 The measure sets out that new content being generated, uploaded, or shared on the service (or which a user seeks to generate, upload, or share) should be analysed before or as soon as reasonably practicable after it can be encountered by United Kingdom users of the service. Relevant content present on the service at the time the technology is implemented should be analysed within a reasonable time.

4.238 The service provider should also regularly obtain the latest version of the list(s) and use them when it analyses content on its service. This will ensure the effective use of these list(s).

4.239 Following content being detected as matching a listed CSAM URL, the measure sets out that service providers should swiftly take down (or prevent from being generated, uploaded, or shared) any such content. The measure leaves discretion to service providers as to how this

[712] IWF. Non-Photographic URL List. [accessed 13 November 2024].

is done – for example, a service provider might only remove the URL in question, or it might choose to take down the whole of the post that contained the URL.

4.240 There are several types of technologies or mechanisms that facilitate direct matching to detect and remove harmful content online, all of which we consider to be effective and meet the requirements of this measure.

4.241 **For detection**, we recommend service providers consider the following automated systems or processes (this list is not exhaustive):

- **'Find and replace' method:** Service providers can use the URL list to 'find and replace' CSAM URL links. This would require an automated system to scan against the URL list to find a direct CSAM URL link match.
- **Hashing the URL list:** Service providers can 'hash' the URL list, which converts the URL links to hashes. They can then use an automated system to scan against the hashes to find a direct CSAM URL link match.
- **Intermediary port:** Service providers can implement an intermediary port between the link that is posted on the service and the destination of the URL link. This facilitates the detection and matching of URLs against the third-party list. In practice, it will be the responsibility of the service provider, rather than the network, to redirect the user to an intermediary port, which then determines whether the user is directed to the third-party URL or is cut off from being redirected to the destination of that URL link. Where a URL link is cut off in such a way, it remains the service provider's responsibility to remove any other instances of the URL, such as 'plain text' displaying the URL destination, from the relevant content.

4.242 **For removal**, we recommend service providers either (1) remove the URL link in a post or (2) the content that contains the CSAM URL. We consider that both options will effectively reduce the harms posed by CSAM (such as user exposure). There are several ways in which providers can achieve these two outcomes:

a) **Removing the URL link in a post:** Service providers should remove the URL link or make it not visible to users on their service, and ensure the links do not direct users to third party destinations containing CSAM. There are several methods for achieving this:

   i) Service providers that adopt an intermediary port and cut off the user (because the link has been detected as a direct CSAM URL link) can redirect users to a splash page that indicates users cannot access the destination of the link. Some URL link providers also provide a splash page for users, which the service provider can adopt if it chooses to use this method. The use of splash pages was recommended by some stakeholders during the consultation and is outlined in paragraph 4.217.[713]
   ii) Service providers can remove a portion of the content that contains a direct URL match. This would visibly alter the content for users on the service.
   iii) Service providers can censor a portion of the content that contains the link, which would produce a blank space or a warning message.
   iv) Service providers can replace the direct URL link match with a broken URL link, but they must also use additional methods to ensure the URL link is no longer visible to users.

[713] Lucy Faithfull Foundation response to November 2023 Consultation, p.5; Barnardo's response to November 2023 Consultation, p.17.

b) **Removing the content:** Alternatively, a service provider may decide to remove the content which contains a direct URL match in its entirety (for example, by removing the entire message or the entire post containing a direct URL link match). Removing the content is a potentially less burdensome and more efficient way to implement URL detection. Content containing CSAM URL links is likely to violate other parts of a provider's terms of service and could amount to illegal content, and therefore removing the content entirely may help to reduce the scale of the harm across the service.

4.243 We discuss these approaches further in the 'Effectiveness' section.

### Securing the URL list(s)

4.244 The measure also sets out that service providers should ensure an appropriate policy is put in place, and security measures taken in accordance with that policy, to secure any copy of a list of CSAM URLs held for the purposes of the measure. This is to protect against unauthorised access, interference or exploitation (for example, the unauthorised addition of URLs or unauthorised disclosure of the list). Such security measures will often be a contractual requirement for access to a list. We slightly strengthened this provision to include the need for a policy to be put in place, to promote good decision-making about which mitigating actions to take. This change was made in response to stakeholder feedback as summarised in paragraphs 4.219. We further detail our response in the 'Costs and risks' section.

4.245 We consider that best practices for mitigating security risks may include, but are not limited to:

- storing data securely within the service's systems;
- restricting access to the CSAM URL list to authorised persons only;
- maintaining records of all authorised persons;
- ensuring all authorised persons have an appropriate understanding of how the measure operates;
- requiring multifactor authentication for access to an account capable of making changes to the CSAM URL list;
- requiring that changes to the CSAM URL list (or how the measure is implemented) are proposed and approved by more than one authorised person;
- retaining records of (1) all changes to the CSAM URL list, (2) all changes to how the measure is implemented, and (3) the authorised person(s) who propose and approve any changes;
- avoiding the use of default or shared passwords and credentials for accounts providing access to the CSAM URL list; and
- ensuring that passwords and credentials are managed, stored, and assigned securely, and are revoked when no longer needed.

## Benefits and effectiveness

### Benefits

4.246 We consider the benefits of URL detection to be significant and, to a large extent, mirror the benefits outlined for the measure on hash-matching CSAM (as set out in paragraphs 4.81 to 4.89).

4.247 CSAM links are shared widely across a range of services, with some service types at particularly high risk of this activity. By sharing links to CSAM, perpetrators can evade hash matching and other forms of detection technology as they do not need to directly share an image on the service.

4.248 The sharing of CSAM links is therefore a significant and growing concern for stakeholders in governments, civil society organisations, and industry.[714] WeProtect Global Alliance highlighted that the increasing risk of link-sharing may be partly due to an increase in the desire for 'on-demand' access to CSAM (rather than curating personal collections), as well as a desire to evade detection through hash matching and content classification.[715] [716] In February 2023, the UK Government launched the Safety Tech Challenge Fund to encourage further innovation in projects disrupting the sharing of links to CSAM, which highlights that this is a growing area of concern.[717]

4.249 Evidence highlights there are some additional harms to those set out relating to the sharing of CSAM images which may occur as a result of the sharing of CSAM URLs.[718] CSAM URLs may not be immediately recognisable as links to CSAM or may be falsely labelled as being links to non-illegal content. As a result, there is the greater potential for a user to click on a link not realising it contains CSAM and cause them to inadvertently view CSAM. Not only can this be distressing to those exposed in this manner, but there is a risk that such accidental viewing of CSAM may, in turn, lead to more regular viewing.[719] Another recent development in online CSAM is the rise of invite child abuse pyramid (ICAP) sites.[720] Evidence indicates that a high volume of links to these sites are being shared across services to generate traffic (and, ultimately, revenue for the site owners). As well as leading to an increase in accidental viewing of CSAM, the revenue generated by this form of link-sharing is likely to be used to perpetrate further harm, including child sexual offences.[721]

4.250 Removing URLs which enable users to encounter CSAM can therefore deliver a number of benefits, including:

[714] Meta (Davis, A.), 2020. Facebook Joins Industry Effort to Fight Child Exploitation Online. [accessed 8 November 2024]; Meta Platforms Ireland response to 2022 Illegal Harms Ofcom Call for Evidence; Goggin, B., Kolodny, L., and Ingram, D., On Musk's Twitter, users looking to sell and trade child sex abuse material are still easily found. NBC News, 6 January 2023. [accessed 8 November 2024].

[715] WeProtect Global Alliance response to November 2023 Consultation, p.15.

[716] Link-sharing and child sexual abuse: understanding the threat - WeProtect Global Alliance [Accessed 18 October 2024].

[717] Gov.UK, 2023. Safety Tech Challenge: link sharing of Child Sexual Abuse Material. [accessed 8 November 2024].

[718] The exception to this is that we note that links to CSAM may include links to non-image based CSAM, such as textual or audio content, which can cause specific and different harms to those caused by image-based CSAM.

[719] Over half (51%) of respondents to a survey of CSAM users on the dark web reported that they had first encountered CSAM accidentally, meaning they were exposed to CSAM without actively searching for it. Source: Insoll, T., Ovaska, O. & Vaarenen-Valkonen, N. 2021. CSAM Users in the Dark Web: Protecting Children Through Prevention. [accessed 18 October 2024]

[720] The IWF defines ICAPs as follows: "These custom-built websites incentivise users to share links to child sexual abuse sites far and wide in a "scattergun" approach, with the aim of recruiting as many 'buyers' as possible. The criminals running the sites benefit from increased web traffic and additional income with offenders potentially buying further videos of child sexual abuse and creating their own links to spam to others."

[721] IWF. Invite Child Abuse Pyramid or ICAP sites | IWF 2023 Annual Report. [accessed 18 October 2024].

- reducing the harm caused by sharing of CSAM to victims and survivors;
- reducing intentional viewing of (or unintentional exposure to) this content;
- reducing subsequent contact sexual abuse; and
- disrupting online sites that seek to generate internet traffic and revenue by sharing links to CSAM.

### Effectiveness

4.251 There is also evidence to demonstrate the effectiveness of removing CSAM URLs from U2U services, which we set out in our November 2023 Consultation[722]. Several stakeholders agreed with our analysis of the measure's effectiveness and, in some cases, provided additional evidence to strengthen our position. As mentioned in paragraph 4.213, in May 2020 the IWF released aggregated data to indicate that at the start of the COVID-19 lockdown in the UK, three ISPs had blocked 8.8million attempts originating from the UK to access webpages that were included in its URL list.[723] It reiterated this in its response to the November 2023 Consultation. It also pointed to an announcement by an IWF member about a sharp increase in the number of attempts to access websites it had blocked in 2023, which the member attributed in part to its use of the IWF's URL list.[724] Additionally, as referenced in paragraph 4.214, WeProtect Global Alliance cited the IWF's URL list as a helpful tool to identify and block access to illegal websites or content. Stakeholders also mentioned 'Project Arachnid' in Canada as an effective technology to combat link-sharing.[725]

4.252 We consider the effectiveness of this measure to rely on (1) the flexibility of technologies that service providers can use to detect and remove CSAM and (2) the accuracy, completeness, and regular deployment of the URL list being used. This measure is designed to support and strengthen the effectiveness of these factors in reducing the spread of (and user exposure to) CSAM.

#### Flexible technology for detection and removal of content

4.253 To maximise effectiveness, we set out provisions that offer service providers some flexibility for implementing the measure. We recommend that direct matching technology is used for URLs, as we consider it to be highly effective in detecting direct matches of known CSAM URLs included on a list. As discussed in the 'How this measure works' section, service providers can use a number of methods to detect and remove CSAM URLs. Our approach does not prescribe specific technologies or mechanisms that providers should use to detect or remove content via direct matching.

4.254 This non-prescriptive approach was supported by some stakeholders, with Stop Scams UK advocating for a dynamic and flexible approach to prescribing specific systems and techniques that service providers should use for moderation and content removal.[726] Meta

[722] Ofcom, 2023. Protecting people from illegal harms online, Volume 4: How to mitigate the risk of illegal harms – the illegal content Codes of Practice, p.89. [accessed 29 November 2024]; Ofcom, 2023. Protecting people from illegal harms online, Annexes 12-16, p. 59. [accessed 29 November 2024].

[723] IWF. Millions of attempts to access child sexual abuse online during lockdown. [accessed 26 November 2024].

[724] IWF response to November 2023 Consultation, p.9.

[725] C3P response to November 2023 Consultation, p.19; WeProtect Global Alliance response to November 2023 Consultation, p.16.

[726] Stop Scams UK response to November 2023 Consultation, p.11.

said that it "recommend[s] avoiding the imposition of a specific technology". Instead, it suggested we support "sharing of best practices" and allow providers to choose from a "variety of solutions" to make use of "the most up to date technologies".[727]

4.255 An example of this is the potential to use a splash page as part of the implementation of the measure, as set out in the 'How this measure works' section. On their website the IWF cite that "[s]ince 2015, splash pages have resulted in 26,000 new users" accessing Stop It Now services via the website.[728]

### Accuracy, completeness, and regular deployment of the URL list

4.256 Direct matching is a well-established, well-understood, and straightforward mechanism for detecting text content on online services. However, the effectiveness of the technology depends on the accuracy of the URL list. The measure includes elements designed to ensure that CSAM URLs are accurately included on the list (see paragraphs 4.234 to 4.236 for a full discussion of the steps taken to mitigate this risk). We consider these provisions will substantially mitigate the risk of content being incorrectly identified as a CSAM URL.

4.257 Adequate sourcing of URL lists will be necessary to ensure the effectiveness of the takedown of content matching listed CSAM URLs. The measure provides for service providers to source one or more appropriate CSAM URL lists from an organisation (or organisations) with expertise in the identification of CSAM. As mentioned above, we have clarified that more than one list can be used, reflecting that (for instance) the IWF maintains a separate URL list for "non-photographic imagery" which is assessed as illegal. This makes clear that it is open to service providers to use multiple lists in combination, which can enable a greater number of URLs for content to be analysed against and provide users with more effective protection.

4.258 As noted above in the hash-matching section of this chapter, we are aware that some service providers use technology to detect and take down content matching other URLs that are not illegal content but nonetheless are prohibited by their terms of service. If a service provider wishes to use a broader list (i.e. one that includes URLs of material other than CSAM) for the purposes of this measure, it will need to ensure that it complies with the measure's provision at least in relation to CSAM URLs in that list. We discuss below in paragraph 4.293 how we expect service providers to deal with complaints from users if content is incorrectly detected as CSAM.

4.259 To ensure the adequacy of the URL lists, the IWF suggested that we formally accredit these lists to ensure their accuracy and quality and recommend their use by service providers.[729] We detail this feedback in paragraph 4.218. However, as outlined in paragraph 4.234, this measure sets out that service providers have the flexibility to select and use any appropriate third-party list of CSAM URLs which meets the requirements set out in the measure. There are several reasons why we do not think it is appropriate to prescribe which list(s) of CSAM URLs should be used:

- We do not have access to information on every active organisation providing lists of CSAM URLs to would allow us to recommend certain lists or organisations over others that may be capable of performing the same function.

[727] Meta response to November 2023 Consultation, annex, p.10.
[728] IWF. URL List. [accessed 26 November 2024].
[729] IWF response to November 2023 Consultation, p.4.

- We want to ensure that the measure is future-proof. Recommending specific organisations or lists could undermine this aim, as the availability and range of organisations providing CSAM URL lists is subject to change over time. This approach would require extensive oversight to ensure quality and effectiveness.

4.260 As previously mentioned, we have designed this measure to support service providers in selecting an adequate URL list. The measure sets out several requirements for service providers to consider with regards to the accuracy and quality of the lists. These elements include ensuring arrangements are in place to (1) secure CSAM URLs are correctly identified, and (2) regularly review listed CSAM URLs and remove any which are no longer CSAM URLs. We detail these requirements in more detail above, in paragraph 4.234 to 4.236. These elements will contribute to the effectiveness of the measure.

4.261 As well as ensuring that URLs are accurately identified as CSAM URLs, it is important that the lists are maintained. The measure sets out that service providers should ensure that arrangements are in place to regularly update the list with identified CSAM URLs, and regularly review listed CSAM URLs and remove any which are no longer CSAM URLs, as well as to secure the list from unauthorised access, interference or exploitation. It also requires the service provider to regularly obtain the latest version of the list(s) for the purpose of detecting content that matches listed URLs. These elements of the measure are designed to ensure its effective implementation, allowing for content matching CSAM URLs to be taken down in a timely way.

4.262 We recognise that relying only on direct matching means this measure is unable to detect URLs that have been altered to evade detection. This may impact the measure's effectiveness in reducing the spread of (and user exposure to) CSAM. While we considered recommending the use of fuzzy matching technology for this measure during the November 2023 Consultation, we ultimately decided to not pursue this option. Several stakeholders noted that fuzzy matching may require increased human moderation, which could increase the burden on providers (including additional costs) as detailed in paragraph 4.221.[730] Protection Group International commented that URL detection is time-consuming for service providers and noted that perpetrators will often alter content to disguise URLs to avoid detection (see paragraph 4.223).[731] Ukie expressed the view that the main costs for URL detection are associated with engineering and labour costs, and for many service providers with low prevalence of CSAM URLs, the value of fuzzy matching is limited (see paragraph 4.216).[732] This feedback supports our decision not to recommend fuzzy URL detection as part of this measure.

4.263 Overall, we consider that our measure will be effective at materially reducing the circulation of CSAM and addressing the harm that we have identified (compared to a counterfactual in which URL detection was not used). Given the significant harm arising from the presence of CSAM online, we consider that the measure will have very significant benefits.

[730] Microsoft response to November 2023 Consultation, p.11; [✂].

[731] Protection Group International response to November 2023 Consultation, p.7.

[732] Ukie response to November 2023 Consultation, p.21.

## Costs and risks

4.264 This section identifies the costs and risks associated with implementing and maintaining a URL detection system as set out in this measure.

### Costs

4.265 Service providers are likely to incur both one-off set up costs developing and implementing the URL detection tool and ongoing costs of maintaining the system and the required software, hardware, and data.

#### One-off costs

4.266 The one-off costs of developing the URL detection tool will primarily be labour costs. The key skillset required will be software engineering, though there may also be involvement from other professional occupation staff. This is consistent with stakeholder feedback received by Ukie in response to our November 2023 Consultation. It noted that the main costs for its members were in relation to engineering and labour costs.[733] We understand that URL detection is complementary to other measures that a service provider may already be implementing. It also complements the perceptual hashing measure set out earlier in this chapter. It is likely that service providers already implementing a complementary measure will be able to carry out URL detection with a small additional cost, primarily in the form of software engineering time.

4.267 We understand that direct matching presents lower implementation costs than other forms of URL detection, such as a fuzzy matching system.

4.268 Our engagement with industry experts during the consultation phase suggests that this measure will be straightforward for small service providers to implement, while providers of large services may require additional resources. A large service is likely to have more complex operational structures, and any changes to the service may involve input from more professionals. The complexity of the service itself will also impact on costs because changes will require more resources where a service has multiple products requiring integration. We estimate it will take approximately around two to 16 months of full-time work for a software engineer to undertake the initial set-up of a CSAM URL detection system. In addition to software engineering time input, initial development and set-up is likely to require a similar amount of time input from a combination of other professional occupation staff. We estimate that these one-off product development costs may range from £20,000 to £300,000.[734] We expect the lower estimate to be more reflective of costs for providers of smaller and less complex services and the upper estimate to be more reflective of costs for providers of larger and more complex services.

#### Ongoing costs

4.269 There will be ongoing labour costs associated with maintaining and updating the system (for example, when integrating new URL lists). Consistent with our standard assumption for the ongoing costs of system changes (as outlined in Annex 5), we assume that annual maintenance costs are likely to be 25% of the initial set-up costs. The annual cost of maintaining and updating a URL detection system is estimated to range from approximately £5,000 to £80,000. In addition, service providers implementing this measure may use one

[733] Ukie response to November 2023 Consultation, p.21.

[734] Details of our assumptions on salaries are included in Annex 5. We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS.

or more URL lists supplied by a third party, which is likely to incur associated costs such as membership fees.[735]

4.270 As mentioned in paragraph 4.221, two stakeholders (Microsoft and [✂] commented that the measure would mean an increase in the need for human moderation.[736] As the measure relies on the direct matching of content to listed URLs, we expect that the content will be detected with a high level of accuracy and with a low risk of false positives. Therefore, we do not anticipate that this measure will incur material additional human moderation costs. However, we acknowledge that, in some cases, users may use the complaints procedure to appeal against the take down of content. Service providers may incur some additional costs from handling these complaints.

4.271 We also acknowledge that there may be other costs to the service provider, as well as costs (both monetary and non-monetary) to other organisations and individuals.[737]

### Risks

4.272 We recognise there are a small number of risks associated with this measure. The deployment of URL detection and removal technologies carries a risk of over-moderating content. For this measure, the specific risks are likely to be:

- a URL being wrongly identified or assessed to be a CSAM URL;
- a URL not being removed from a URL list once CSAM is no longer present; and
- unauthorised access and/or changes to the URL list leads to URLs being wrongly added (or removed) from the list.

4.273 The design of the measure seeks to maximise the accuracy of the URL lists and reduce the risk of over-moderation. These safeguards require service providers to ensure that the third-party list provider has arrangements in place to:

- ensure that suspected CSAM URLs are correctly identified before they are added to the list(s);
- regularly update the list or lists with identified CSAM URLs;
- regularly review listed CSAM URLs and remove any URLs that are no longer CSAM URLs (with the aim of ensuring that URLs do not remain on the list for longer than is needed; and
- secure the list(s) from unauthorised access, interference, or exploitation (to ensure they are not altered by unauthorised individuals to wrongly include URLs or remove CSAM URLs).

4.274 In most instances, we would expect the URL included on the list to be the URL of the specific webpage at which CSAM is hosted (rather than the whole domain) to avoid 'over-

[735] To provide an example, the IWF currently provide a URL list, in addition to other services including an image hash database. The IWF's membership fees to support its work can range from £1,000 to over £90,000 per year, based on the industry sector and size of the company. The membership list available online demonstrates that smaller services are accessing this membership at the lower end of this cost. Source: IWF, 2024. Our Members. [accessed 22 October 2024].

[736] Microsoft response to November 2023 Consultation, p.11; [✂].

[737] We note that potential costs associated with the measure to the broader hash-matching ecosystem were highlighted by some stakeholders in their responses to the November 2023 Consultation responses (see paragraph 4.222).

blocking' of legitimate content. We emphasise that CSAM in this context is not limited to indecent and prohibited images. This measure also applies to URLs that include content giving advice about grooming or abusing a child sexually, or obscene articles encouraging the commission of CSEA offences. This includes URLs that include content linking to (or otherwise directing users to) CSAM, or that advertise the distribution or showing of indecent or prohibited images. For example, a link to a webpage which included links to indecent images would be appropriate to include on the URL list (even if that webpage also included legitimate content).

4.275 We consider it appropriate to list at domain level where a domain is entirely or predominantly dedicated to CSAM. This is likely to be more effective and efficient than listing each individual URL containing CSAM, as these may alter frequently. That said, service providers should ensure that the provider of their URL list(s) has arrangements in place to ensure that listing at domain level only occurs in such cases. While we recognise that, where a list includes a domain that is predominantly dedicated to CSAM, URLs at that domain which do not themselves include CSAM would be affected, we consider this to be proportionate given the clear risk that users accessing those URLs will go on to encounter CSEA content on other pages at that domain.

4.276 As mentioned in paragraph 4.221, some stakeholders suggested an increase in human moderation resources would be required to mitigate the risk of content being wrongly removed from services.[738] In particular, in paragraph 4.221, we outlined that Microsoft argued that the current state of technology for CSAM URL matching is insufficient to enable implementation of the measure without the risk of over-moderation. It argued that the measure could result in a significant number of moderation decisions that would later be overturned on appeal, noting that this would likely require a large increase in the need for human moderation to investigate for further context.[739]

4.277 Ultimately, we do not consider there to be a significant risk of over-moderation as a result of this measure given it recommends direct matching rather than fuzzy matching, which means only duplicates of pre-identified CSAM can be detected and removed. As we expect the risk of over-moderation to be minimal (if any), this measure does not set out requirements for human moderation. However, we do expect the implementation of this measure will require human oversight to update the URL list supplied by a third party. The implementation of the measure may also lead to an increase in appeals if users believe their content has been wrongfully taken down, and the review of such appeals may require more human resources.

4.278 We recognise that the URL list itself may be vulnerable to security compromises, including unauthorised access, interference and exploitation. This risk was reiterated by stakeholders responding to the November 2023 Consultation as outlined in paragraph 4.219.[740] The simpler the implementation of the URL detection technology, the higher the risk of the service being attacked by perpetrators to gain access to the URLs in question. In the November 2023 Consultation, we proposed that URL lists should be secured from security

[738] Microsoft response to November 2023 Consultation, p.11; [✂].

[739] Microsoft response to November 2023 Consultation, p.11.

[740] Glitch response to November 2023 Consultation, p.9; Global Partners Digital response to November 2023 Consultation, pp.14-15; INVIVIA response to November 2023 Consultation, p.15; Meta response to November 2023 Consultation, annex, p.10; Microsoft response to November 2023 Consultation, pp.13-14; Proton response to November 2023 Consultation, p.6.

compromises and that appropriate measures should be taken to secure any copy of the list held by or for the service. Having considered stakeholder responses commenting on the need to secure the list, we have made a minor adjustment to the measure to strengthen general security standards by requiring service providers put in place a policy for its security measures to mitigate the risks of unauthorised access. We consider that, with this adjustment, the safeguards we have specified adequately mitigate the risk of security compromises.

4.279 We also recognise there is a risk of bias in the compilation of URL lists, as raised by stakeholders in the consultation (paragraph 4.220).[741] The addition of URLs to a list depends on where the content is found online, how it is detected (through AI/ML models, web crawling, or human analysts), and the subsequent assessment of content as CSAM (for example, age determination and categorisation). This may create biases that underrepresent the scale and nature of the problem of CSAM for victims and survivors of different ages or who belong to minority groups. We consider that this is likely to be mitigated by the elements of the measure which promote the accuracy and effectiveness of the list (see paragraph 4.231 (b) and also the 'Effectiveness' section). However, we have responded to these concerns by revising the measure to add that service providers should ensure that the arrangements for identifying or assessing suspected CSAM URLs do not plainly discriminate on the basis of protected characteristics (within the meaning of Part 2 of the Equality Act 2010), such as race or sex.

4.280 Overall, we consider that the way we have designed this measure includes suitable mitigations for the risks we have identified and that these risks can be managed. As discussed in this section, risk levels can be reduced and mitigated through the practices of both service providers implementing the technology and organisations which compile and maintain URL lists. We have therefore designed the measure with a number of safeguards detailed in this chapter and adjusted some of the safeguards following our stakeholder feedback to our consultation. We assess the measure's impacts on rights in the next section.

## Rights impact

4.281 This section considers the measure's impacts on users' rights under Articles 8 and 10 of the ECHR.

4.282 As explained in 'Introduction, our duties, and navigating the Statement', as well as in chapter 14 of this Volume: 'Statutory tests', Article 10 of the ECHR sets out the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 8 of the ECHR sets out the right to respect for individuals' private and family life.

4.283 In essence, restrictions on those rights must be necessary and proportionate – that is, the measure's contribution to its objective must outweigh its adverse impacts.

4.284 Our assessment of the adverse impacts of the measure is therefore to be balanced against the measure's contribution to its objective of reducing harm associated with the dissemination of CSAM through posting links and URLs on U2U services.

[741] OSTIA response to November 2023 Consultation, p.14; Glitch response to November 2023 Consultation, p.8.

4.285 Parliament has legislated for CSAM to be designated as "priority illegal content" under the Act and imposed a number of duties on service providers to protect users from harm from such content. These include requiring service providers to take measures to prevent individuals from encountering priority illegal content by means of the service and to implement 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 the prevalence and dissemination of CSAM online, relating to each of the prevention of crime, the protection of health and morals, and the protection of the rights of others. The contribution of removing links and URLs which disseminate CSAM to these legitimate aims is essentially the same as that of detecting, removing, and reporting CSAM, which was described in the rights impact section for the hash-matching measure. The 'benefits and effectiveness' section discusses the measure's effectiveness and the associated benefits in more detail.

### Freedom of expression

4.286 An interference with the right to freedom of expression must be prescribed by law, pursue a legitimate aim, be proportionate to the legitimate aim and correspond to a pressing social need.

4.287 Interference with users' freedom of expression arises principally where links and URLs are incorrectly detected as CSAM URLs and taken down by the systems and processes implemented in accordance with the measure. This could also affect website/database providers' freedom of expression (so far as it reduces traffic to the website or database in question). In most cases, taking down links and URLs that are CSAM URLs will not engage Article 10 ECHR at all (and, in this respect, we note that such links and URLs will often themselves be "priority illegal content" for the purposes of the Act, for instance where posting the link is done to encourage or assist the commission of the offence of viewing an indecent image of a child). Where the link was shared without such intent, such as in outrage or disgust, we consider that taking down the link is clearly proportionate to the legitimate aims outlined above, given the serious harm that disseminating CSAM causes.

4.288 We assess the risk of links and URLs being incorrectly detected as CSAM URLs to be very low. While Big Brother Watch said that automated content moderation systems often result in the removal of lawful content (as explained in the 'Risks' section), we consider that deploying URL matching technology to detect direct matches with URLs on a list will be highly accurate.[742] Whether matched content is a CSAM URL will therefore depend largely on the accuracy of the URL list. The design of the measure includes a number of elements intended to ensure that the list used is accurate which, in turn, operate as safeguards to protect users' freedom of expression:

- It sets out the need for the service provider to ensure that the person (or persons) from whom it has sourced the list (or lists) has expertise in the identification of CSAM and has arrangements to be in place to: (a) secure that CSAM URLs are correctly identified before being added to the list; (b) regularly review CSAM URLs on the list and remove any which are no longer CSAM URLs; and (c) secure the list from unauthorised access, interference or exploitation.

[742] Big Brother Watch response to November 2023 Consultation, p.8.

- It sets out that the service provider should regularly obtain the latest version of any list (or lists) and use it for analysing content to assess whether it consists of, or includes, content matching a listed CSAM URL.

4.289 The measure also sets out that service providers should have an appropriate policy in place to secure any copy of a list (or lists) held for the purposes of the measure from unauthorised access, interference, or exploitation, and that measures are taken in accordance with that policy.

4.290 We acknowledge, however, that there could be a small number of errors. This could include cases where a URL is wrongly assessed as being a CSAM URL, or where CSAM has been removed from a URL added to the list but there is a time lag before the person compiling the list reviews it and removes it, or before the service provider obtains the revised list.

4.291 Interference with users' freedom of expression may also arise where a URL includes legitimate content as well as CSAM, such as (in particular) where a domain has been added to the list due to the content at the domain predominantly comprising CSAM or content related to CSEA content.[743] However, we consider that listing at domain level in such a case is justified (given the risk that users accessing the domain will go on to encounter CSAM).

4.292 The measure also specifies other Codes measures as safeguards for users' freedom of expression including, in particular (1) those enabling users to complain if their content has been taken down on the basis that it is illegal content, and (2) those ensuring that service providers' terms of service give information about the proactive technology used and the relevant policies and processes for complaints. These Codes measures help to safeguard users' freedom of expression in a number of different ways, including in providing a level of transparency for users about the technology used and how to make a complaint.

4.293 Interference with users' freedom of expression may also arise where service providers take subsequent action (for example, a user's account being banned) against users wrongly detected as sharing CSAM as a result of an incorrectly identified CSAM URL. As with the hash-matching measure, the Code of Practice does not, at this stage, include a measure about such action, and the action to be taken would be a matter for the service provider. Such action could, however, have more significant impacts than the take-down of content and it would be important for service providers to have regard to those impacts when deciding on their safety policies.

4.294 However, as explained in the hash-matching measure (paragraph 4.153), the safeguards included in this measure to protect users' freedom of expression would also help to limit the risk that action is taken against users on the basis of an incorrectly identified CSAM URL. Service providers are also required to enable users to make complaints if the provider has given a warning to the user, suspended or banned the user from the service, or in any other way restricted the user's ability to use the service, as a result of content shared by the user which the provider considers to be illegal content.

4.295 Overall, while we acknowledge the measure involves some interference with users' right to freedom of expression where URLs are incorrectly actioned as a CSAM URL, or where a URL contains legitimate content as well as CSAM, we consider that interference to be small, appropriately limited with the safeguards for freedom of expression we have in place, and

[743] This point could also affect website/data providers' freedom of expression.

proportionate to the measure's aim of reducing users' exposure to CSAM on other services.[744]

### Privacy

#### Data protection

4.296 The processing of personal data for the purposes of the measure by (or on behalf of) a service provider should be limited to the automated analysis of content to detect whether it includes content matching a listed CSAM URL. While automated processing can lead to data protection harms, service providers will need to ensure that the automated processing, and other associated processing, is carried out in accordance with data protection law.

4.297 The measure involves the assessment of content at suspected CSAM URLs by organisations providing URL lists, which will involve the processing of personal data (as in, information which relates to an identified or identifiable person) of victims and survivors and others. Such work is already undertaken by various child protection organisations and law enforcement authorities, but our measure could result in additional processing taking place. We expect these organisations to have robust security and to ensure that any processing is carried out in accordance with data protection law, which will safeguard against privacy risks. Overall, victim and survivor rights will be safeguarded by the measure because it will help reduce access to CSAM depicting them.

4.298 The measure also provides for both the third-party list provider(s) and the service provider to secure the URL list from unauthorised access, interference, or exploitation, which operate as safeguards to the privacy right of any individuals which may be identifiable in the content contained at, or via, the URLs. As described in paragraph 4.245, this includes that the service provider should have an appropriate policy in place to secure any copy of a list (or lists) held for the purposes of the measure from unauthorised access, interference, or exploitation.

4.299 We are satisfied that the processing required by the measure can be carried out in accordance with data protection law. Service providers should refer to the ICO's guidance on content moderation and data protection which explains how data protection law applies to content moderation technologies and processes and provides advice to help service providers comply with the UK GDPR and the Data Protection Act 2018.[745] This includes advising service providers to carry out a data protection impact assessment to assess and mitigate data processing risks.

#### The right to privacy under Article 8 ECHR

4.300 We now consider the impact on users' right to privacy under Article 8 ECHR more broadly.

4.301 An interference with the right to privacy must be in accordance with the law, pursue a legitimate aim, be proportionate to the legitimate aim and correspond to a pressing social need.

4.302 Where the automated processing involved in the measure is carried out in compliance with data protection law, that processing should have a minimal impact on users' privacy. The degree of interference also depends to a degree on the extent to which the nature of the

[744] We have also taken into consideration the impacts on website/database providers.

[745] ICO, 2024. Content moderation and data protection. [accessed 18 October 2024].

affected content and communications is public or private or, in other words, gives rise to a legitimate expectation of privacy. As explained in paragraph 4.5, the measure only applies in relation to content communicated publicly by means of the service. This means that the user sharing the content should generally have a reduced expectation of privacy in connection with that content (compared to, for instance, content shared privately with a small number of other persons).[746]

4.303 Where CSAM is detected, service providers may (in certain circumstances) be required (or may choose) to report the relevant user to a law enforcement authority (or to a designated reporting body such as NCMEC). Relevantly, section 66 of the Act (which is not yet in force) sets out duties for providers of regulated U2U services to report to the NCA detected CSEA content that is not otherwise reported, which may result in information about the user also being provided.[747]

4.304 So far as users are correctly reported for posting CSEA content, any interference with their rights to privacy is prescribed by the relevant legislation and, in enacting the legislation, Parliament has already made a judgement that such interference is a proportionate way of securing the relevant public interest objectives. However, in cases where a link has been incorrectly detected as a CSAM URL, this reporting would constitute a significant interference with the affected user's privacy.

4.305 As with the hash matching measure (paragraph 4.176), we have adopted the ICO's suggestion to specify other code measures as safeguards for users' privacy.[748] We consider that the same measures that act as safeguards for users' freedom of expression are relevant here, 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.

4.306 Interference with users' privacy could result from action taken against users by service providers based on an incorrectly identified CSAM URL. Such impacts could be significant (for instance, they could affect a user's ability to participate in economic, social, cultural and leisure activities), and could be especially serious if it were in some way possible for other persons to infer why that action had been taken. At this stage, the Code of Practice does not provide for such action to be taken, which would be a matter for the service provider. It will be important for service 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.

4.307 Overall, we consider that the impact of the measure on users' rights under Articles 8 and 10 of the ECHR are proportionate to the measure's aim of reducing users' exposure to CSAM on other services.

[746] The ICO argued that if our guidance on content communicated publicly and privately did not provide "sufficient direction and certainty", there would be "a risk that some services will default to assessing content as being communicated publicly. This would undermine the effectiveness of the privacy safeguard in practice". ICO response to November 2023 Consultation, p.22. We address this in Volume 3, Chapter 4: 'Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act'.

[747] In the case of a non-UK provider, the duty is limited to "UK-linked" CSEA content: s.66(2).

[748] ICO response to November 2023 Consultation, pp.12-13, 16-17.

## Who this measure applies to

4.308 In the November 2023 Consultation we set out our provisional view that the URL detection measure we had proposed should apply to the following services:

- large services which are at medium or high risk of CSAM URLs in their risk assessment;
- other services which are at high risk of CSAM URLs in their risk assessment and have more than 700,000 monthly UK users.

4.309 Following the consultation, we have decided to proceed with the approach we proposed. We conclude that our approach is proportionate considering the scale and severity of CSAM online, our analysis of the effectiveness of the measure, the costs to service providers of implementing it, and its impact on user rights.

4.310 We considered several factors when deciding the scope of service providers this measure would apply to, including: (1) the risk presented by CSAM to users and other individuals, (2) the severity and nature of the harm associated with the dissemination of CSAM, and (3) the differing size and capacity of service providers.

4.311 Considering the severity of the harm and the significant benefits associated with removing CSAM URLs, we consider it proportionate for large services which are medium or high risk for CSAM URLs and other services which are high risk for CSAM URLs and which have more than 700,000 monthly active United Kingdom users to incur the costs of the measure. By applying the measure to these service providers, we are prioritising reducing the spread of CSAM URLs on services with the largest numbers of users and service providers with the greatest capacity to implement this measure. As explained below, we have decided not to broaden the scope of service providers that the measure applies to beyond what was proposed in the November 2023 Consultation. This is mainly due to uncertainties about the costs to smaller service providers and the capacity of the CSAM URL list providers. Given these uncertainties, we favour a phased approach to broadening the scope of the measure, and we will review our decision in the future.

4.312 We have also decided not to extend the measure to apply to smaller services which are medium risk for CSAM URLs. At this stage, we do not consider it would be proportionate to apply the measure to providers of such services given the resource requirements for URL detection.

4.313 However, given the severity of harm caused by CSAM, we consider it proportionate to apply this measure to providers of smaller services which are high risk for CSAM URLs. We have therefore decided to apply the measure to all services which are high-risk for CSAM URLs and which have 700,000 or more monthly UK users. We have included this user threshold to mitigate the possible risk of overwhelming third-party providers of databases for relevant URLs in light of the limited availability of URL databases that service providers can access to implement this measure. As such, we are adopting a phased approach to applying this measure to providers of high-risk services with more than 700,000 monthly UK users and will reconsider the scope in future iterations of the codes.

### How service providers should assess their risk level for CSAM URLs

4.314 As part of this Statement, we have published Risk Assessment Guidance for Service Providers. Part 3 of that guidance includes guidance to help providers of U2U services to make an assessment of the risk for CSAM URLs. As set out in Table 13.2 of the guidance, we would normally expect a service to be assessed as **high risk** where:

- there is evidence that CSAM URLs have been shared to a significant extent on the service; or
- the service's functionalities allow users to share text or hyperlinks without creating a user account.

4.315 We would normally expect a service to be assess as **medium risk** where:

- there is evidence that CSAM URLs have been shared on the service, but not to a significant extent; or
- the service's functionalities enable users to share text or hyperlinks and several of the specific risk factors specified in the guidance have been identified.[749]

4.316 Service providers should refer to Table 13.2 of the guidance for more information, including as to when it could be appropriate to assess at a lower risk level.

4.317 We consider that this measure is proportionate for the services to which we have decided to apply it.

### File-sharing and file storage service providers

4.318 Although the factors that we have considered when deciding on the scope of the measure are similar to those regarding the scope of the hash-matching measure (as outlined in the 'Who this measure applies to' section) above, the scope of service providers is different. In particular, this measure does not treat file-storage or file-sharing service providers differently to providers of other types of services.

4.319 Unlike in the case of image-based CSAM, we do not have clear evidence that file-storage and file-sharing services generally pose a higher risk for the dissemination of CSAM URLs than other types of service.

4.320 We are aware of some evidence that suggests file-storage and file-sharing services may also enable the sharing of CSAM, as perpetrators can distribute URLs directing users to these collections.[750] We are continuing to build this evidence base to better understand the risks posed by and prevalence of CSAM URLs on file-storage and file-sharing service providers.

### Broadening the scope of service providers

4.321 We also received some feedback from stakeholders, in response to the November 2023 Consultation, that the scope of the measure is too narrow, arguing that providers of smaller services should be required to implement the measure (see paragraphs 4.226).[751] Some of the reasons given by stakeholders to justify broadening the scope of this measure include:

- a broader scope would do more to reduce harm,

[749] The specific risk factors specified in Table 13.2 of the guidance are: (a) child users; (b) social media services; (c) messaging services; (d) discussion forums and chat rooms; (e) user groups; (f) direct messaging; (g) encrypted messaging.

[750] WeProtect, 2021. Global Threat Assessment 2021. [accessed 13 November 2024].

[751] Barnardo's response to November 2023 Consultation, pp.13, 16; C3P response to November 2023 Consultation, pp.4, 12; Children's Commissioner response to November 2023 Consultation, p.19; IWF response to November 2023 Consultation, p.31; Marie Collins Foundation response to November 2023 Consultation, p.11; [✂]; NSPCC response to November 2023 Consultation, p.24; The Cyber Helpline response to November 2023 Consultation, p.10.

- proportionality concerns should not prevent the measure from effectively being implemented across all services, as small platforms with limited reach still carry the risk of harm,
- there are free tools available that allow smaller service providers to scan their services for CSAM URLs, and
- there are specific harms to children that could be reduced by including smaller services within scope of the measure.

4.322 We acknowledge these points, and we recognise that the measure does not apply to all services which carry CSAM URLs. However, we need to ensure that the measure is proportionate, including consideration of the costs to service providers of implementing the measure. We recognise that the technology is available for free to some service providers, but we note that technology costs are only one of several kinds of costs that providers will incur when implementing the measure, as detailed in the 'costs and risks' section. We have not received sufficient evidence on the magnitude of these other types of costs to broaden the scope of the measure at this stage.

4.323 Another reason to favour a phased approach to broadening the scope of this measure is uncertainty about whether third-party database providers can meet the additional demand due to the measure. This is because there are a very limited number of organisations that provide adequate CSAM URL lists; fewer than the number of organisations providing image-based CSAM hash lists. As such, we are interested to see how this market responds to additional demand due to the measure. For now, the scope of the measure remains the same as in the November 2023 Consultation.

4.324 Although the measure does not apply to high or medium risk services which do not meet the relevant user thresholds, we note that other relevant measures may apply to these services, including the measures relating to content moderation, governance processes and tracking evidence of new and increasing illegal harm.[752]

## Conclusion

4.325 We consider this measure to be an effective means to reduce the risk of users on U2U services encountering CSAM on a service, with substantial benefits.

4.326 Having considered the costs, risks and associated impacts on (in particular) the rights of users, we consider it to be a proportionate safety measure to recommend providers of in-scope U2U services take for the purpose of compliance with their illegal content safety duties (in particular, the duty under section 10(2)(a) to use systems and processes designed to prevent individuals encountering priority illegal content).

4.327 We have also designed the measure to incorporate a number of safeguards for the protection of the rights of users to freedom of expression and privacy.

4.328 We have therefore decided to include the measure in the CSEA Code of Practice. It is referred to as ICU C10.

[752] See governance measures ICU A2, ICU A3, ICU A5, ICU A6, and ICU A7 and content moderation measures ICU C1-C8. (Some of these measures only apply to large services and/or multi-risk services.)

# Keyword detection relating to articles for use in fraud

4.329 In the November 2023 Consultation, we proposed that large U2U services with a high or medium risk for fraud should use standard keyword detection technology to detect content which is likely to amount to a priority offence concerning articles for use in frauds.[753] This proposal only applied in relation to content communicated publicly on U2U services and where it would be technically feasible to analyse that content.

4.330 Evidence indicated that very specific keywords tend to be used by criminals to offer to supply articles for use in frauds (for example, stolen personal and financial credentials). Our provisional view was that standard keyword detection technology could be an effective means to proactively identify content likely to amount to an offence concerning articles for use in frauds at pace and at scale.

4.331 In our November 2023 Consultation, we recognised that more advanced keyword search, detection, or filtering methods may already be in use by some services, such as involving machine learning or artificial intelligence. However, there was a limited evidence base on the accuracy of these newer technologies and, therefore, we did not propose their use in the proposed measures.

## Summary of stakeholder responses

4.332 A number of stakeholders, representing law enforcement, consumer protection organisations, and various industries expressed varying degrees of support for this measure, or for the idea of keyword detection being part of a solution to address content relating to articles for use in frauds.[754]

4.333 Several stakeholders commented on and, in some cases, expressed concerns about the proposed measure including:

- the effectiveness of the measure;
- the general approach used to address articles for use in frauds;
- the costs and risks associated with implementing this measure; and
- the impacts on users' rights.

4.334 We outline this feedback in the following section.[755]

[753] Schedule 7 of the Act provides that a number of offences concerning articles for use in frauds should be considered as priority offences. These include the offence of making or supplying of articles for use in frauds (including offers to supply these) under section 7 of the Fraud Act 2006, and related inchoate offences.

[754] Association of British Insurers (ABI) response to November 2023 Consultation, p.1; BILETA response to November 2023 Consultation, p.11; Cifas response to November 2023 Illegal Harms Consultation, p.8.; [✂]; CELE response to November 2023 Consultation, p.9; Innovate Finance response to November 2023 Illegal Harms Consultation, p.6; Match Group response to November 2023 Consultation, p.10; Monzo response to November 2023 Illegal Harms Consultation, pp.11-13; National Trading Standards (NTS) eCrime team response to November 2023 Illegal Harms Consultation, p.8; [✂].

[755] We have summarised responses where a number of stakeholders had similar views.

## Feedback on the measure’s effectiveness

4.335 We received several responses from stakeholders on the availability (and effectiveness) of more sophisticated tools, compared to keyword detection.[756]

4.336 Several stakeholders highlighted the deficiencies of keyword detection technology as a standalone tool, noting that keyword detection is not the most effective means to identify content relating to articles for use in frauds. Which?, Cifas, Innovate Finance, and OSTIA explained that keyword detection is outdated and frequently manipulated by criminals who can adapt their terminology to circumvent detection.[757] TechUK, Google, and Trustpilot explained that services use more sophisticated measures to tackle fraud.[758]

4.337 We received a number of suggestions that standard keyword detection technology is not sufficiently effective at tackling the harm, either due to high volumes of false positives[759] or because of the need to utilise the technology in combination with other tools and signals.[760] A stakeholder noted that keyword detection will not be effective in isolation because it is a simplistic approach to fraud prevention.[761] Another stakeholder noted that manually curating a keyword list as a means of combating fraud is not a measure that scales, and includes a significant risk of generating false positives.[762] A stakeholder suggested that while standard keyword detection can be useful in some contexts, it has limitations over time (such as excessive false positives) that mean it would not be effective at tackling this harm if used in isolation.[763]

4.338 Although some stakeholders suggested that the proposed measure would be insufficient as a standalone tool, stakeholders did not provide evidence to assert that the technology is entirely unusable for the purpose of detecting the content in question. Conversely, some stakeholders indicated that keyword detection is part of the solution, but it may not be effective in isolation. They explained that a tailored approach would be more effective.[764]

4.339 A substantial number of stakeholders, from a variety of industries, pointed out the limits to fraud keyword detection systems, indicating that our proposal risked lowering the bar by disincentivising innovation and investment in safety technology. They were concerned that

[756] Airbnb response to November 2023 Consultation, p.15; Booking.com response to November 2023 Consultation, p.19; Google response to November 2023 Consultation, p.43; Integrity Institute response to November 2023 Consultation, p.11; Meta response to November 2023 Consultation, annex, pp.7, 11; Monzo response to November 2023 Consultation, pp.5, 13; OSTIA response to November 2023 Consultation, p.12; Reddit response to November 2023 Consultation, pp.9, 12; Revolut response to November 2023 Consultation, pp.14-15; Stop Scams UK response to November 2023 Consultation, p.11; techUK response to November 2023 Consultation, pp.17, 25; Trustpilot response to November 2023 Illegal Harms Consultation, pp.23-24; Which? response to November 2023 Illegal Harms Consultation, p.10.

[757] Cifas response to November 2023 Consultation, pp.9, 10; Innovate Finance response to November 2023 Consultation, pp.6-11; OSTIA response to November 2023 Consultation, p.12; Which? response to November 2023 Consultation, p.10.

[758] Google response to November 2023 Consultation, pp.43; techUK response to November 2023 Consultation, p.17; Trustpilot response to November 2023 Consultation, p.24.

[759] [✂]; [✂].

[760] Innovate Finance response to November 2023 Consultation, p.7; Integrity Institute response to November 2023 Consultation, p.11.

[761] [✂].

[762] Reddit response to November 2023 Consultation, pp.9, 10, 12, 23.

[763] Meta response to November 2023 Consultation, pp.7, 11.

[764] Innovate Finance response to November 2023 Consultation, pp.6-11; Integrity Institute response to November 2023 Consultation, p.11; Revolut response to November 2023 Consultation, pp.14-15.

services may instead choose to do the bare minimum to benefit from the safe harbour provided by the Codes.[765] A number of stakeholders proposed that we take a technology-agnostic and outcomes-focused approach.[766]

4.340 Some stakeholders highlighted examples of alternative measures that are already in use or would be beneficial for services to implement. These included:

- URL detection;[767]
- Image detection;[768]
- Video detection;[769]
- Machine learning (classifiers);[770]
- Artificial intelligence;[771]
- Red flag indicators;[772] and
- Meta analysis, including behavioural, data, technical signals and automated pattern analysis.[773]

### Source of the keyword list

4.341 Some stakeholders raised concerns regarding the source of keyword lists. Concerns included whether in-scope services can have ongoing and up-to-date access to authoritative sources of information for fraud keywords;[774] how keyword lists would be updated and adapt to account for shifting criminal tactics,[775] and whether service providers using different lists would cause challenges in identifying fraudulent content.[776] One stakeholder noted that challenges exist regarding how appropriate keyword search terms could be sourced and updated.[777] Another stakeholder suggested use of AI or word pattern analysis to identify keywords in real time.[778]

[765] Airbnb response to November 2023 Consultation, p.15; Booking.com response to November 2023 Consultation, p.19; Cifas response to November 2023 Consultation, p.10; Google response to November 2023 Consultation, p.43; Innovate Finance response to November 2023 Consultation, pp.6-11; Monzo response to November 2023 Consultation, pp.5, 13; OSTIA response to November 2023 Consultation, p.12; Stop Scams UK response to November 2023 Consultation, p.11.

[766] Cifas response to November 2023 Consultation, p.10; Google response to November 2023 Consultation, p.43; Innovate Finance response to November 2023 Consultation, p.11; Monzo response to November 2023 Consultation, p.14.

[767] Financial Conduct Authority (FCA) response to November 2023 Illegal Harms Consultation, p.9; UK Finance response to November 2023 Consultation, p.16; Which? response to November 2023 Consultation, p.5-6.

[768] ABI response to November 2023 Consultation, p.2; UK Finance response to November 2023 Consultation, pp.2, 13.

[769] ABI response to November 2023 Consultation, p.2.

[770] FCA response to November 2023 Consultation, p.9; Google response to November 2023 Consultation, pp.43; [✂]; Trustpilot response to November 2023 Consultation, p.24; UK Finance response to November 2023 Consultation, pp.2, 13; Which? response to November 2023 Consultation, p.10.

[771] Lloyds Banking Group response to November 2023 Illegal Harms Consultation, p.5; Trustpilot response to November 2023 Consultation, p.24; UK Finance response to November 2023 Consultation, pp.2, 3; Which? response to November 2023 Consultation, p.10.

[772] Cifas response to November 2023 Consultation, p.10.

[773] Reddit response to November 2023 Consultation, pp.10, 23; UK Finance response to November 2023 Consultation, pp.2, 13; Integrity Institute response to November 2023 Consultation, p.11.

[774] Microsoft response to November 2023 Consultation, p.11.

[775] Cifas response to November 2023 Consultation, p.10; Lloyds Banking Group response to November 2023 Consultation, pp.5, 9; Meta response to November 2023 Consultation, annex, p.11.

[776] Lloyds Banking Group response to November 2023 Consultation, p.9.

[777] [✂].

[778] [✂].

4.342 We provisionally noted in the November 2023 Consultation that service providers can develop lists in-house or obtain them from third parties as long as they undergo bespoke testing to ensure they are appropriate and effective for their service. Stakeholders endorsed our view that service providers will need to adapt the lists to the nature of their service. One stakeholder explained that they have deployed internal tools such as keyword detection at scale, with robust lists of key terms for detection that expand as their awareness of new and emerging harms grows. They also noted that publicly available lists may not be suitable for all services.[779] Another stakeholder explained that it uses its own keyword detection tool to detect keywords that may be associated with frauds on its service. It reviews keywords as new patterns emerge.[780]

### Six-month review period

4.343 A small number of stakeholders challenged the proposed recommendation that service providers should update keyword lists at a minimum every six months,[781] and the FCA noted that it could risk issues not being addressed early enough.[782] Some noted the importance of keywords being detected and developed in real-time.[783] Cifas explained that the success of keyword detection tools will be dependent on live and active knowledge of the terms being used by the criminal community.[784]

## Feedback on the measure's costs and risk

4.344 Multiple stakeholders expressed concern over the costs associated with implementing keyword detection tools. The feedback suggested that there would be a wide variety of costs in relation to engineering, labour, and maintenance, and that these costs may be disproportionate, as a big portion of the costs would be associated with the need for human input in the moderation process, and may result in diverting resources away from potentially more impactful measures.[785]

4.345 Conversely, a number of stakeholders (including those from the banking sector, the public sector, and an industry representative) suggested that the costs would be proportionate, and they would not be a barrier for large service providers.[786] Banking sector stakeholders highlighted the cost impact on the financial services sector with regard to fraud.[787]

4.346 OneID provided more general feedback on the costs of automated content moderation, noting that the costs of computing power are lowering over time, and the availability of

[779] [✂].

[780] [✂].

[781] Cifas response to November 2023 Consultation, 2023, p.10; FCA response to November 2023 Consultation, p.8; Revolut response to November 2023 Consultation, p.15.

[782] FCA response to November 2023 Consultation, p.8.

[783] [✂]; [✂].

[784] Cifas response to November 2023 Consultation, p.10.

[785] [✂]; [✂]; [✂]; Name Withheld 3 response to November 2023 Consultation, p.12; Roblox response to November 2023 Consultation, pp.21-22; Ukie response to November 2023 Consultation, p.21.

[786] ACT The APP association response to November 2023 Consultation, p.12; Innovate Finance response to November 2023 Consultation, p.13; Monzo response to November 2023 Consultation, p.15; National Trading Standards eCrime team response to November 2023 Consultation, p.10; OneID response to November 2023 Illegal Harms Consultation, p.2.

[787] Innovate Finance response to November 2023 Consultation, p.13; Lloyds Banking Group response to November 2023 Consultation, p.6; Monzo response to November 2023 Consultation, pp.12, 15; Revolut response to November 2023 Consultation, p.17; UK Finance response to November 2023 Consultation, p.15.

new tools such as AI means that service providers should be able to build in robust content-checking processes.[788] [789]

## Feedback on the proposed approach

### Application of the measure at offence level

4.347 A few stakeholders, particularly service providers, suggested that the measure would be disproportionate for large services who are at a medium or high risk of fraud if their risk is not associated with articles for use in frauds.[790] A service provider suggested that the measure should be targeted at services that are at genuinely high risk of carrying relevant illegal content.[791]

### Expand the measure to cover other kinds of illegal harms

4.348 A small number of stakeholders criticised the measure for only applying to fraud and recommended that the measure be expanded to cover other harms.[792] Snap noted that keyword detection should be used more broadly to support a service's efforts to detect and moderate illegal content, not only fraud, and in particular, articles for use in frauds.[793] The Molly Rose Foundation noted that services already operate keyword lists to detect harmful content relating to suicide and self-harm.[794] The IWF and UKSIC recommended the application of this measure to cover CSAM content.[795]

### Expand the measure to cover other fraud types

4.349 A number of service providers, financial services stakeholders, and an anti-fraud organisation commented on our proposal to focus the measure on content likely to amount to an offence concerning articles for use in frauds (rather than fraud and financial services offences more generally).[796] [✂] was broadly supportive of keyword detection, with the caveat that this should not be the only measure deployed in fraud prevention.[797] One stakeholder suggested that keyword detection can be useful in the context of illegal promotions and investment frauds, as long as there are safeguards in place.[798] However, the views expressed by some stakeholders were consistent with our proposal to not expand the measure to cover other fraud types, specifically investment scams. They agreed that standard keyword detection is not necessarily the best suited automated content moderation tool for the detection of all types of fraud, given that more advanced tools are in use.[799]

[788] Computing power refers to the capability of a computer system to perform tasks and process data.

[789] OneID response to November 2023 Consultation, p.2.

[790] Booking.com response to November 2023 Consultation, p.18; Roblox response to November 2023 Consultation, pp.21-22; Snap response to November 2023 Consultation, p.13.

[791] [✂].

[792] IWF response to November 2023 Consultation, p.32; Molly Rose Foundation response to November 2023 Consultation, p.37; Name withheld 4 response to November 2023 Consultation, p.5.

[793] Snap response to November 2023 Consultation, p.12.

[794] Molly Rose Foundation response to November 2023 Consultation, p.37.

[795] IWF response to November 2023 Consultation, p.32; UKSIC response to November 2023 Consultation, p.39.

[796] Cifas response to November 2023 Consultation, p.9; Innovate Finance response to November 2023 Consultation, p.8; Monzo response to November 2023 Consultation, pp.11, 12; UK Finance response to November 2023 Consultation, p.16.

[797] [✂].

[798] [✂].

[799] Airbnb response to November 2023 Consultation, p.15; [✂]; Monzo response to November 2023 Consultation, p.13; [✂].

4.350 Some, however, particularly banking sector stakeholders, expressed concern about the measure being too narrow in scope and suggested a wider measure should be considered which focuses on more than just identifying content relating to articles for use in frauds and which is not limited to keyword detection. They suggested that such a measure could help to detect purchase scams and other types of fraud.[800] A banking stakeholder also emphasised the importance of tackling investment scams, not only articles for use in frauds.[801]

### Expand the measure to cover other services

4.351 Some stakeholders suggested that the application of the measure should be expanded to cover more services.[802] Innovate Finance recommended exploring the merits of expanding the measure to cover SMEs.[803] [✂].[804]

4.352 One stakeholder provided a general comment on the impact of automated content moderation on SMEs, noting that recommendations for SMEs to implement automated content moderation tools may have unintended consequences and could lead to substantial resource demands (both financial and human).[805]

4.353 The IWF suggested that the measure be expanded to also cover search services.[806]

## Feedback on the measure's impact on users' rights

4.354 A number of industry stakeholders raised concerns regarding potential impacts on user rights, highlighting the risk of over-enforcement, bias and/or false positives. Responses suggested that the use of keyword detection could result in over-moderation and that the removal of legitimate content may have a disproportionate downstream impact on certain user groups in particular.[807]

4.355 One stakeholder made general comments about the potential impact of proactive technology on users' rights.[808] Two stakeholders highlighted the importance of safeguards when utilising automated content moderation tools, ranging from the use of human moderators and sampling detected content to identify false positives.[809]

4.356 A number of stakeholders highlighted the potential impact on rights and user privacy that could result from applying content moderation tools to private communications.[810]

[800] Innovate Finance response to November 2023 Consultation, p.8; [✂]; UK Finance response to November 2023 Consultation, p.16.

[801] [✂].

[802] Innovate Finance response to November 2023 Consultation, p.13; Which? response to November 2023 Consultation, pp.2-3.

[803] Innovate Finance response to November 2023 Consultation, p.13.

[804] [✂].

[805] [✂].

[806] IWF response to November 2023 Consultation, p.12.

[807] Are, C. response to November 2023 Consultation, p.8; Big Brother Watch response to November 2023 Consultation, pp.6-7; Integrity Institute response to November 2023 Consultation, p.11; Microsoft response to November 2023 Consultation, p.11; [✂]; Reddit response to November 2023 Consultation, pp.9-10.

[808] CELE response to November 2023 Consultation, p 9.

[809] CELE response to November 2023 Consultation, p.9; Ukie response to November 2023 Consultation, pp.19-20.

[810] [✂]; Big Brother Watch response to November 2023 Consultation, pp.6-7; ICO response to November 2023 Consultation, p.22; Meta response to November 2023 Illegal Harms Consultation, p.20.

4.357 The Molly Rose Foundation said that our focus on the possible impacts on freedom of expression triggered by the potential for keyword detection to generate a high volume of false positives should not outweigh the merits of the measure as a meaningful way to reduce exposure to harm. This view was provided in relation to the use of keyword detection by services when tackling suicide and self-harm content.[811]

4.358 We also received input from the Integrity Institute that the measure presupposes the availability of keywords in multiple language to reduce the risks of biases.[812]

## Our decision

4.359 In view of the November 2023 Consultation responses, we have decided not to proceed with this measure now. Instead, we are currently considering evidence surrounding the use of automated tools to proactively detect illegal content (which would deal with a range of harmful content), going beyond the automated detection measures we have already consulted on. We intend to consult on this in Spring 2025. Given the concerns raised about the efficacy of our keyword detection proposal, we think it is better to focus on this other analysis than devote further time to implementing a keyword detection measure at this stage.

4.360 This does not affect our decision to include code measures recommending the use of hash-matching and URL detection for CSAM, which we consider will make an important contribution to efforts to combat CSAM. These measures are by their nature focused on preventing the re-upload and circulation of CSAM which has already been identified and hashed/added to a URL list. We recognise that they do not address harm from 'novel' CSAM, as well as other forms of CSEA (such as grooming activity). Our work on additional measures to proactively detect illegal content could therefore help to close this gap and could be an important complement to our existing CSAM measures, and an important element of our strategy for addressing CSAM.

4.361 In the meantime, we want to thank the large number of stakeholders, including anti-fraud experts, service providers and public sector bodies that have taken the time to engage with this proposal and supplement our evidence base.

[811] Molly Rose Foundation response to November 2023 Consultation, p.37.

[812] Integrity Institute response to November 2023 Consultation, p.11.

# 5. Automated search moderation

## What is this chapter about?

Search services use automated moderation tools to identify large volumes of harmful content more quickly, and these are therefore critical to many services' attempts to reduce harm. This chapter sets out our recommendation of a measure for services to take steps to remove URLs identified as hosting child sexual abuse material ('CSAM') from search results, why we are recommending it, and to which search services it should apply.

## What decisions have we made?

We are recommending the following measure:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| ICS C7 | Providers should take action to ensure that users do not encounter, in or via search results, search content present at or sourced from URLs on a list of URLs previously identified as hosting CSAM. | Providers of general search services. |

## Why have we made this decision?

The circulation of CSAM online is increasing rapidly. The evidence presented in the Register of Risk shows that perpetrators often use search services to access CSAM. As we explained above, child sexual abuse and the circulation of CSAM online causes significant and lifelong harm and the ongoing circulation of this imagery can re-traumatise victims and survivors of sexual abuse. URL detection is an effective and well-established tool for combatting the circulation of CSAM on search services. The largest search services are already using it to address CSAM. Whilst the use of URL detection imposes some costs we consider these are justified given the severity of the harm they address and the significant benefits of limiting exposure to known CSAM.

## Introduction

5.1 Given the volume of content on websites or databases that may be searched by search engines, we expect automated systems and processes to play an important role in search service providers' compliance with their illegal content safety duties. This is particularly the case in relation to service providers' duty to use proportionate systems and processes designed to minimise the risk of individuals encountering search content that is priority illegal content.[813]

5.2 In this chapter, we explain our decision to include a measure in the CSEA Code of Practice ('Code') relating to the automated moderation of search content. The measure outlines the

[813] Section 27(3)(a) of the Online Safety Act.

steps we recommend all providers of general search services should take to prevent users from encountering (in or via search results) child sexual abuse material (CSAM) at URLs identified as hosting, or being part of a website dedicated to, this kind of priority illegal content. [814] [815]

# Measure on removing listed CSAM URLs from search results

5.3 In our November 2023 Illegal Harms Consultation ('November 2023 Consultation'), we proposed a measure recommending that providers of general search services deindex URLs at which CSAM is present, or which include a domain which is entirely or predominantly dedicated to CSAM.[816]

5.4 We proposed that these service providers should source an appropriate list of CSAM URLs from a third party that (1) has expertise in the identification of CSAM and (2) meets other criteria specified in the measure. This list would need to be regularly monitored to identify newly added CSAM URLs and service providers would need to take steps to deindex URLs added to the list and reinstate to the search index those removed from the list.

5.5 We explained that the aim of this measure was to reduce users' exposure to CSAM in search results. The online circulation of CSAM causes serious and potentially lifelong harm by re-traumatising victims and survivors of sexual abuse. We considered that removing these URLs using automated systems was an effective and well-established means of addressing this online harm and would be proportionate to recommend to providers of general search services, and that the measure would assist service providers in complying with the illegal content safety duties under section 27(2) and (3) of the Online Safety Act ('the Act').

## Summary of stakeholder responses[817]

5.6 A range of stakeholders, including providers of regulated services, governments and law enforcement, academics, and civil society organisations, expressed broad support for our proposed measure.[818]

[814] See our 'Overview of regulated services' chapter for further information about general search services. For other definitions, including 'URLs' see Annex 3: Glossary.

[815] *'CSAM' refers to indecent or prohibited images of children, or other material which contains advice about grooming or abusing a child sexually or which is an obscene article encouraging the commission of other child sexual exploitation and abuse offences. It also includes content which links or otherwise directs users to such material, or which advertises the distribution or showing of CSAM. CSAM is priority illegal content under the Act.*

[816] See Annex 3: Glossary.

[817] This summary is not an exhaustive list of stakeholder responses, and further responses can be found in Annex 1.

[818] Betting and Gaming Council response to November 2023 Illegal Harms Consultation, p.9; Canadian Centre for Child Protection (C3P) response to November 2023 Illegal Harms Consultation, p.21; Centro de Estudios en Libertad de Expresion y Acceso a la Informacion (CELE) response to November 2023 Illegal Harms Consultation, p.10; Children's Commissioner for England response to November 2023 Illegal Harms Consultation, p.22; Dwyer, D. response to November 2023 Illegal Harms Consultation, p.9; Internet Watch Foundation (IWF) response to November 2023 Illegal Harms Consultation, p.8; INVIVIA response to November 2023 Illegal Harms Consultation, p.18; Mencap response to November 2023 Illegal Harms Consultation, p.11; [✂]; National Society for the Prevention of Cruelty to Children (NSPCC) response to November 2023 Illegal Harms

5.7 Many of these stakeholders generally supported the use of automated systems to remove priority illegal content from search results.[819] In particular, one stakeholder supported its use citing that all service providers that are high risk or medium risk for image-based CSAM, regardless of size, need to implement automated content moderation otherwise there is a risk of creating spaces for people to disseminate CSAM freely without detection.[820]

5.8 A number of respondents also agreed that the measure reflected industry best practices and/or would be effective at detecting, removing and reducing the spread of CSAM online.[821] One stakeholder suggested that automated moderation systems and processes are effective at reducing the spread of illegal content and urged Ofcom to collaborate with stakeholders to establish global common standards for the use of this tooling.[822]

5.9 However, several stakeholders commented on or raised concerns about specific elements of the proposed measure and/or suggested changes to the measure.[823] This feedback related to:

- references to "deindexing";
- the benefits and effectiveness of the proposed measure, including factors impacting its efficacy;
- the costs and risks associated with implementing the proposed measure;
- the impacts on users' rights; and
- who the measure applies to.

## Feedback on deindexing approach and terminology

5.10 Similarly to feedback provided in response to our search moderation measures, Google suggested that the specific recommendation of "deindexing" could be inappropriate.[824] [825] Google commented on how it understands the term 'deindexing' to differ from 'delisting' content from the search index. It described how it operates one index for all its country

Consultation, p.30; Nexus response to November 2023 Illegal Harms Consultation, p.13; Philippine Survivor Network response to November 2023 Illegal Harms Consultation, p.10; Segregated Payments Ltd response to November 2023 Illegal Harms Consultation, p.10; South East Fermanagh Foundation response to November 2023 Illegal Harms Consultation, p.13; The Cyber Helpline response to November 2023 Illegal Harms Consultation, p.15; Welsh Government response to November 2023 Illegal Harms Consultation, p.4; WeProtect Global Alliance response to November 2023 Illegal Harms Consultation, p.17.

[819] Betting and Gaming Council response to November 2023 Consultation, p.9; INVIVIA response to November 2023 Consultation, p.18.

[820] Marie Collins Foundation response to November 2023 Illegal Harms Consultation, p.11.

[821] Segregated Payments Ltd response to November 2023 Consultation, p.10; techUK response to November 2023 Illegal Harms Consultation, p.26.

[822] techUK response to November 2023 Consultation, p.26.

[823] 5Rights Foundation response to November 2023 Illegal Harms Consultation, p.21; C3P response to November 2023 Consultation, p.21; Cybersafe Scotland response to November 2023 Illegal Harms Consultation, p.10; Glitch response to November 2023 Illegal Harms Consultation, p.9; Google response to November 2023 Illegal Harms Consultation, p.47; Marie Collins Foundation response to November 2023 Consultation, p.11; [✂]; Protection Group International response to November 2023 Illegal Harms Consultation, p.9; South East Fermanagh Foundation response to November 2023 Consultation, p.13; Welsh Government response to November 2023 Consultation, p.4; WeProtect Global Alliance response to November 2023 Consultation, p.17; Yoti response to November 2023 Illegal Harms Consultation, p.14.

[824] See chapter 3 of this Volume: 'Search moderation', paragraphs 3.29-3.32 and 3.57-5.59 for Google's feedback and our subsequent response.

[825] Google response to November 2023 Consultation, p.37.

services and suggested that “a deletion from the index would have the result of a global takedown” which it said could go beyond the policy intent of the Act. While it said that it did deindex (rather than delist) CSAM, in other cases it argued references to deindexing should be replaced to provide the service provider with “flexibility to decide whether to deindex or delist”.[826]

5.11 Google also argued that to specifically recommend deindexing of URLs would make it “technically overly burdensome” for service providers to reinstate this content when appropriate (such as when a URL is removed from an externally-sourced CSAM URL list).[827] It also suggested this could be inconsistent with the actions it should take as a result of appeals and complaints (noting that, unlike deindexing, “if content has been blocked from serving through delisting, it can be reinstated to results immediately”).[828]

5.12 Google further argued that the proposed measure should not stipulate that URLs must be reinstated once removed from an externally-sourced CSAM URL list. It asserted that “service providers should have agency to determine reinstatements, particularly in cases where violative material (for CSAM or other legal/policy reasons) remains on the page”, and suggested that we clarify that service providers “have agency to not re-insert content in the search index if it violates their content guidelines or terms of service.”[829] We address these concerns in the ‘How this measure works’ section.

### Feedback on benefits and effectiveness

5.13 Some stakeholders expressed concerns about the effectiveness of the measure or suggested ways in which it could be enhanced.[830] techUK referred to the role that “global multi-stakeholder programmes and initiatives” could play in addressing online harm and urged us to create “common standards that are globally scalable to protect global users consistently”.[831] In the context of our specific proposed measure, this feedback is highlighting the risk of service providers taking inconsistent approaches to sourcing of URL lists in the absence of central coordination to ensure an appropriate standard. Protection Group International (‘PGI’) raised a related concern, arguing that “unless there is a coordinated approach, then this is an impossible request for companies to achieve”.[832] While PGI said that the proposed measure was good “in theory”, it raised a number of questions about which organisations would provide lists of URLs and if the proposed measure was practical.[833] We address these concerns in the ‘Benefits and effectiveness’ section.

[826] Google response to November 2023 Consultation, p.37.
[827] Google response to November 2023 Consultation, p.47.
[828] Google response to November 2023 Consultation, p.37.
[829] Google response to November 2023 Consultation, p.47.
[830] 5Rights Foundation response to November 2023 Consultation, p.21; C3P response to November 2023 Consultation, p.21; Cybersafe Scotland response to November 2023 Consultation, p.10; Protection Group International response to November 2023 Consultation, p.9; techUK response to November 2023 Consultation, p.26; Welsh Government response to November 2023 Consultation, p.4; Yoti response to November 2023 Consultation, p.14.
[831] techUK response to November 2023 Consultation, p.26.
[832] Protection Group International response to November 2023 Consultation, p.9.
[833] Protection Group International response to November 2023 Consultation, p.9.

5.14 Similarly, some stakeholders noted the importance of having robust processes for monitoring and updating the list of URLs.[834] PGI highlighted the need for providers of lists to regularly review the URL databases or lists to ensure the included content is up to date (ensuring, for example, that URLs that have had CSAM removed are removed from the lists), and for service providers to make use of that updated list (including reinstating URLs that have been removed from the list).[835] We address this concern primarily in the 'Benefits and effectiveness' section, and it is also referenced in the 'How this measure works' section.

5.15 Some stakeholders expressed concerns regarding the scope of the measure. The Canadian Centre for Child Protection (C3P) said that it was important that the measure also addressed cached pages (copies of page content at a different URL). [✂].[836] Cybersafe Scotland expressed concern that the proposed measure would not address livestreaming of child sexual abuse, arguing that the majority of children it works with experience harm from CSAM from livestreams.[837] We address this concern in the 'Benefits and effectiveness' section.

## Feedback on costs and risks

5.16 Some stakeholders highlighted several risks that they felt might affect the implementation, application, and efficacy of the measure.[838]

5.17 Glitch highlighted the security vulnerabilities of URL lists and the need to strengthen safeguards to protect against exploitation of them and/or unauthorised access to them. It stated that the measure addresses security vulnerabilities of URL lists but does not discuss how these vulnerabilities may disproportionately impact women and girls. It further asked how we will account for potential exploitation of the security and privacy protections set out in the measure.[839] We address this concern in the 'Costs and risks' section in the 'Risk of continued removal of search content'.

5.18 We also received stakeholder input regarding the security vulnerabilities of databases and/or URL lists in response to our proposals on automated content moderation for user-to-user ('U2U') service providers.[840] Proton raised the security concerns associated to the proposed automated moderation technology.[841] We consider this feedback to also be relevant to this measure. We address these concerns in the 'How this measure' works' section (specifically 'Securing the URL list(s)').

5.19 In response to our proposals on automated content moderation for U2U services, but also of relevance to this measure, we received input from the Online Safety Tech Industry Association (OSTIA) and Glitch noting a potential risk relating to biases in hash databases.

[834] Protection Group International response to November 2023 Consultation, p.9; Welsh Government response to November 2023 Consultation, p.4.

[835] Protection Group International response to November 2023 Consultation, p.9.

[836] [✂].

[837] Cybersafe Scotland response to November 2023 Consultation, p.10.

[838] Glitch response to November 2023 Consultation, p.9; [✂]; Microsoft response to November 2023 Illegal Harms Consultation, pp.13-14; Online Safety Tech Industry Association (OSTIA) response to November 2023 Illegal Harms Consultation, pp.10-16; Proton response to November 2023 Illegal Harms Consultation, p.6.

[839] Glitch response to November 2023 Consultation, p.9.

[840] Global Partners Digital response to November 2023 Illegal Harms Consultation, p.14; INVIVIA response to November 2023 Consultation, p.16; Microsoft response to November 2023 Consultation, pp.13-14; Proton response to November 2023 Consultation, p.6.

[841] Proton response to November 2023 Consultation, p.6.

They said these may disproportionately (mis)represent certain groups and alter perpetrator behaviour.[842] For example, if perpetrators determine that a certain type of CSAM is not adequately captured in a hash database, they may target this content to avoid detection. In its response, OSTIA suggested we explicitly recommend that databases should avoid "systematic bias within their control", stating that databases should "determine addition of content solely based on whether or not it is CSAM and ensure minimisation of bias in processes making that determination".[843] We address these concerns in the 'Costs and risks' section (specifically 'Potential biases in hash databases').

### Feedback on users' rights

5.20 We received stakeholder feedback reiterating the need to protect users' rights and offering suggestions for strengthening the measure's safeguards to protect these rights. In particular, Glitch argued that website operators whose URLs have been deindexed should be notified, stating that "failure to notify may disproportionately impact women and girls, who may rely on their online platforms for livelihoods or advocacy efforts."[844] We address this concern in both the 'Costs and risks' and the 'Rights impact' sections.

### Feedback on who the measure applies to

5.21 Several stakeholders provided feedback regarding the services the measure should apply to. Specifically, [✂], South East Fermanagh Foundation[845], and Marie Collins Foundation proposed that all search services, regardless of size, should use automated tools to deindex CSAM URLs.[846] Reasons included:

- improved consistency in the protection of children;
- avoidance of bias due to human moderation; and
- ensuring that smaller platforms not considered medium or high risk for CSAM cannot be used by perpetrators without detection.

5.22 We also received input from WeProtect Global Alliance questioning why vertical search services were not included within the scope of the proposed measure.[847]

5.23 We address these points and concerns in the 'Who this measure applies to' section.

## Our decision

5.24 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation. We have made a small number of minor amendments to the measure in response to the feedback set out in the 'Summary of stakeholder responses' section. These changes, developed in more detail in the following sections, include:

[842] Glitch response to November 2023 Consultation, p.8; OSTIA response to November 2023 Consultation, p.14.

[843] OSTIA response to November 2023 Consultation, p.14.

[844] Glitch response to November 2023 Consultation, p.9.

[845] To note, the South East Fermanagh Foundation response pertained to terrorism and proscribed organisations. However, we determined the feedback was also relevant to the detection and removal of CSAM.

[846] Marie Collins Foundation response to November 2023 Consultation, p.11; South East Fermanagh Foundation response to November 2023 Consultation, p.13; [✂].

[847] WeProtect Global Alliance response to November 2023 Consultation, p.17.

- Redrafting the measure to replace references to "deindexing" with a more flexible approach which recommends service providers take action to ensure that United Kingdom users of the service do not encounter, in or via search results, search content that is present at or sourced from listed URLs or URLs that contain a listed domain. We explain some examples of the types of search content this includes at paragraph 5.33.
- Clarifying that service providers may use more than one URL list.
- Addressing potential risks relating to bias, by providing that service providers should ensure that the arrangements in place for identifying and assessing suspected CSAM URLs for potential inclusion on the list do not plainly discriminate on the basis of protected characteristics (such as sex or race).
- Specifying that service providers should ensure that an appropriate policy is in place, and measures are taken in accordance with that policy, to secure URL lists from unauthorised access, interference or exploitation.
- Amending the measure to state that action taken in relation to a listed URL or listed domain should be swiftly reversed once the URL or domain has been removed from the list, unless the service provider considers that this would be inappropriate (for example, because of other illegal content still present at the URL).
- Specifying other Codes measures which act as safeguards for users' rights to freedom of expression, including allowing "interested persons"[848] to appeal against the removal of their content from search results.

5.25 In brief, our measure sets out that providers of all general search services should take action to ensure that users do not encounter, in or via search results, search content present at or sourced from URLs on a list of URLs previously identified as hosting CSAM.

5.26 The measure can be found in full in our Illegal Content Codes of Practice for search services, within which we refer to this measure as ICS C7. It forms part of the CSEA Code of Practice.

## Our reasoning

### How this measure works

5.27 This measure sets out that all providers of general search services should take action to ensure that United Kingdom users of the service do not encounter, in or via search results, search content that is present at or sourced from listed URLs or URLs that contain a listed domain. For these purposes, it sets out that service providers should source one or more lists of CSAM URLs from a person (or persons) with expertise in identifying CSAM and who meet requirements designed to ensure the list of URLs is accurate and effectively maintained (as explained further below).

[848] An "interested person" is a person that is responsible for a website or database capable of being searched by the search engine in question, where (in the case of an individual) the individual is in the United Kingdom or (in the case of an entity) the entity is incorporated or formed under the law of any part of the United Kingdom). See section 227(7) of the Act.

5.28 A 'CSAM URL' is defined as a URL at which CSAM is present,[849] or a domain which is entirely or predominantly dedicated to CSAM.[850] In most cases, we would expect lists to be at URL (not domain) level, so that illegal content can be specifically targeted. However, we recognise that, where a list includes a domain that is predominantly dedicated to CSAM, URLs at that domain which do not themselves include CSAM would be affected. We consider this to be proportionate given the clear risk that users accessing those URLs will go on to encounter CSEA content on other pages at that domain.[851]

### Selection of a URL list

5.29 As mentioned above, the measure sets out that service providers should source one or more lists of CSAM URLs from a person (or persons) with expertise in identifying CSAM.[852] We have clarified that the provider may use more than one list, reflecting that using more than one list can better protect users.[853]

5.30 The measure sets out requirements that should be met for a list to be appropriate to use for the measure. These include:

  - The organisation from which the list is sourced has arrangements in place to identify URLs or domains suspected to be CSAM URLs, and secure (so far as possible) that they are correctly identified as CSAM URLs before being added to the list.

- An additional requirement that these arrangements for identifying or assessing suspected CSAM URLs do not plainly discriminate on the basis of protected characteristics, such as sex or race.[854] This change has been made in response to comments from stakeholders about the potential risk of bias in databases or lists of CSAM (for example, if CSAM relating to boys or girls were to be systematically excluded.

5.31 The measure also includes requirements to ensure that the list is effectively maintained, and its integrity assured. The organisation from which the list is sourced is required to have arrangements in place to:

- regularly update the list with identified CSAM URLs;
- regularly review listed CSAM URLs and remove from the list any which are no longer CSAM URLs (i.e. where the CSAM at the URL has been taken down); and
- secure the list from unauthorised access, interference or exploitation.

[849] As explained in our Illegal Content Judgements Guidance, a link to CSAM should usually itself be considered CSAM, and therefore a URL containing one or more links to CSAM would usually be a CSAM URL for the purposes of this measure.

[850] The measure provides that a domain is "entirely or predominantly dedicated to CSAM" if the content present at the domain, taken overall, entirely or predominantly contains CSAM (such as indecent images of children) or content related to CSEA content.

[851] Notwithstanding Google's view that "any removal request" recommended in our Codes should be "at the URL level to avoid the risk of over-removal" and that "domain-based actions should be limited to "downranking" or "demotions" within search results. Google response to November 2023 Consultation, p.38.

[852] We refer to such persons as an organisation for ease of reference.

[853] For example, the IWF maintains a main URL list and a separate URL list for non-photographic imagery which is illegal in the UK but may not be illegal in other jurisdictions.

[854] The "protected characteristics" (as specified in Part 2 of the Equality Act 2010) are age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex and sexual orientation.

### Detection and removal of URLs

5.32 **The service provider should use the list or lists it has sourced to take action to ensure that United Kingdom users of the service do not encounter, in or via search results, search content that is present at or sourced from listed URLs or URLs that contain a listed domain.** We have redrafted this part of the measure to replace references to 'deindexing' with a more flexible approach, reflecting that our concern is with the outcome (and in response to Google's comments summarised above at paragraphs 5.10 to 5.12).

5.33 In particular, this means that the search service should not return search results which link to the listed URL or a URL which contains a listed domain (or provide a search result for the webpage giving its URL but not as a hyperlink). Search content 'sourced' from a listed CSAM URL should also not be returned. This would include:

- an extract of text present at the listed CSAM URL presented to users in search results associated with that URL or domain;
- an image or video present at the listed CSAM URL presented to users in search results associated with that URL or domain, such as in image search functionality; and
- a cached version of the listed CSAM URL created as an archive of that URL or domain.[855]

5.34 However, the measure concerns search content associated with the listed URL or domain. It does not recommend service providers to identify other cases where the same content is associated with another URL (for instance, where the same image present at a listed URL is on another webpage searched by the search engine).

5.35 The service provider should also ensure that, when a previously listed URL or listed domain is removed from the list, the action it has taken is swiftly reversed, unless the provider considers that this would be inappropriate. This is important to ensure that search content is not removed from search results without cause, and to limit interference with rights to freedom of expression. However, we have amended the measure to make clear that service providers need not reinstate search content where this would be inappropriate (for example, because other illegal content continues to be present at the URL in question). This change was made in response to Google's comments summarised at paragraph 5.10 to 5.12. The measure provides that service providers should regularly monitor the lists for the purpose of removing and reinstating search content related to listed URLs and URLs that contain a listed domain. We expect that this process will be automated (save that there may be human involvement in decisions as to whether to reinstate search content).

### Securing the URL list(s)

5.36 The measure also sets out that service providers should ensure an appropriate policy is put in place, and security measures taken in accordance with that policy, to secure any copy of a list of CSAM URLs held for the purposes of the measure. This is to protect against unauthorised access, interference or exploitation (for example, the unauthorised addition of URLs or unauthorised disclosure of the list). Such security measures will often be a contractual requirement for access to a list. We slightly strengthened this provision to include the need for a policy to be put in place, to promote good decision-making about which security measures to take. This change was made in response to stakeholder

[855] This explanation responds to feedback from the Canadian Centre for Child Protection (C3P) (see paragraph 5.15).

feedback as summarised in paragraphs 5.17 and 5.18. We further detail our response in the 'Costs and risks section'.

5.37 We consider that best practices for mitigating security risks may include, but are not limited to:

- storing data securely within the service's systems;
- restricting access to the CSAM URL list to authorised persons only;
- maintaining records of all authorised persons;
- ensuring all authorised persons have an appropriate understanding of how the measure operates;
- requiring multifactor authentication for access to an account capable of making changes to the CSAM URL list;
- requiring that changes to the CSAM URL list (or how the measure is implemented) are proposed and approved by more than one authorised person;
- retaining records of (1) all changes to the CSAM URL list, (2) all changes to how the measure is implemented, and (3) the authorised person(s) who propose and approve any changes;
- avoiding the use of default or shared passwords and credentials for accounts providing access to the CSAM URL list; and
- ensuring that passwords and credentials are managed, stored, and assigned securely, and are revoked when no longer needed.

## Benefits and effectiveness

### Benefits

5.38 As set out in our Register of Risks ('Register') chapters titled 'CSEA' and 'Search', both exposure to CSAM – and the ability to access it – causes serious harm. The evidence highlights that search services play an important role in enabling perpetrators to find CSAM online. Indeed, search services are one of the most common means used by individuals to find CSAM. For example, a small sample study of 20 men who had accessed indecent images of children online found that 65% of participants used search services to access CSAM.[856] This, in turn, perpetuates harm as access to CSAM contributes to and is associated with a variety of child sexual abuse offences. For example, one study found that 42% of self-reported perpetrators who had viewed CSAM online went on to seek direct contact with a child afterwards.[857]

5.39 Another study showed that efforts to remove CSAM URLs by two large search services resulted in a 67% reduction in CSAM related queries between 2013 and 2014 in the United

[856] Bailey, A., Allen, L., Stevens, E., Dervley, R., Findlater, D., and Wafers, S. Pathways and prevention for indecent images of children offending: A qualitative study, Sexual Offending: Theory, Research, and Prevention, 17. [accessed 30 October 2024].

[857] Insoll, T., Katariina Ovaska, A., Nurmi, J, Aaltonen, M. and Vaaranen-Valkonen, N., 2022. Risk Factors for Child Sexual Abuse Material Users Contacting Children Online: Results of an Anonymous Multilingual Survey on the Dark Web, Journal of Online Trust & Safety, 1 (2). [accessed 21 October 2024]

States (compared to another service which undertook no such efforts and saw no corresponding decrease).[858]

5.40 While we recognise this approach cannot eliminate the risk of encountering CSAM content via search services, the removal of known CSAM URLs contributes to the overall minimisation of the risk that users encounter CSAM by means of a search service. We have determined that the use of automated moderation tooling to reduce the spread of CSAM on search services is proven to be effective.

5.41 We consider removing search results which enable users to encounter CSAM can therefore deliver a number of benefits including:

- reducing the harm caused to victims and survivors by the sharing of CSAM;
- reducing both intentional viewing of and unintentional exposure to CSAM; and
- reducing subsequent contact sexual abuse.

### Effectiveness

5.42 The effectiveness of this measure in ensuring that users do not encounter CSAM in or via search results (and, in turn, securing the benefits described above), depends on the URL list(s) used and the implementation of the measure by the service provider.

5.43 We assume that service providers will implement the measure using automated processes which can ensure that all search content relating to a particular CSAM URL is removed. We therefore focus below on the URL lists.

5.44 The identification and removal of CSAM URLs relies on service providers having access to adequate URL lists. We understand that there are lists available to service providers that can support this measure. For example, the Internet Watch Foundation ('IWF') provides a list of webpages containing child sexual abuse images and videos to companies who want to block or filter them for their users' protection. This list is updated by the IWF twice a day, removing and adding URLs.[859] In 2023, the IWF's URL list contained on average 8,351 URLs at any one time, with a total of 194,580 unique URLs included on the list at some point over the course of the year.[860] [861] A further 295 unique URLs of non-photographic CSAM were listed on the IWF's Non-Photographic Imagery List as at the end of 2023.[862] An average of 1,116 new URLs were added to the URL list each day, demonstrating the dynamic nature of such lists, as well as the importance of ensuring that lists are regularly updated and monitored by service providers to ensure accuracy.

5.45 We consider the effectiveness of this measure relies on the quality of the URL lists, which depends on several factors including, but not limited to:

- adequate sourcing of the URL lists;

[858] Steel, C.M.S., 2015. Web-based child pornography: The global impact of deterrence efforts and its consumption on mobile platforms. [accessed November 26 2024].

[859] IWF, 2023. IWF URL List. [accessed 11 November 2024].

[860] IWF, 2023. IWF URL List. [accessed 11 November 2024].

[861] URLs are added to this list while processes are underway to have the content removed from the service, to ensure that no users are exposed to the content of the URL during the investigation and actioning process. Once the content has been removed, the URL is removed from the list. As such, the URLs on the list are changing as actioned content is removed and newly-discovered URLs are added.

[862] IWF Non-photographic child sexual abuse, IWF 2023 Annual Report.

- the accuracy and type of content included in the database; and
- effective maintenance of the URL lists.

5.46 These factors will require appropriate consideration by service providers. To support this, the measure sets out requirements for a URL list to be deemed appropriate to use by a service provider. We consider that setting out these factors in the measure will support its effectiveness in removing CSAM URLs from search results.

### Adequate sourcing of the URL lists

5.47 Adequate sourcing of URL lists will be necessary to ensure the effectiveness of the measure. This was reiterated by several stakeholders in response to the November 2023 Consultation, as detailed in paragraphs 5.13, who outlined concerns about adequate sourcing of lists.[863] Some stakeholders also referred to the importance of setting minimum standards, and collaboration with international stakeholders, for the selection of URL lists.[864] We recognise the value of minimum standards for, and central coordination of, databases and lists to ensure optimal quality and performance. As a result, we have designed the measure to communicate standards for service providers to select an appropriate list.

5.48 To support the sourcing of adequate URL lists, the measure sets out requirements for what makes a list of CSAM URLs appropriate to use for the measure and how it should be used by service providers to identify and remove content. The measure provides for service providers to source one or more appropriate CSAM URL lists from an organisation with expertise in the identification of CSAM. As mentioned above, we have clarified that more than one list can be used, reflecting that (for instance) the IWF maintains a separate URL list for "non-photographic imagery" which is assessed as illegal. This makes clear that service providers may use multiple lists in combination, which can enable search results for a greater number of URLs to be removed and so provide users with more effective protection.

5.49 We consider that this approach should ensure service providers use appropriate URL lists and are not convinced that a more prescriptive approach (such as one in which we designate URL lists for use by providers) is needed. We explain our reasons for this in chapter 4 of this Volume: 'Automated content moderation' when discussing the measure on detecting CSAM URLs on U2U services.

5.50 In chapter 4 of this Volume: 'Automated content moderation', we noted that some service providers use technology to detect and take down content matching URLs that are not illegal content but nonetheless are prohibited by their terms of service. Similarly, some search service providers may use URL lists to identify search content that is not illegal but contravenes their content policies. If a service provider wishes to use a broader list (i.e. one that includes URLs of material other than CSAM) for the purposes of this measure, it will need to ensure that it complies with the measure's provision at least in relation to CSAM URLs in that list. We discuss our expectations about complaints in the 'Risks' sub-section (see paragraph 5.76).

[863] Protection Group International response to November 2023 Consultation, p.9; Welsh Government response to November 2023 Consultation, p.4.

[864] 5Rights Foundation response to November 2023 Consultation, p.22; techUK response to November 2023 Consultation, p.26.

### Accuracy and type of content included in the URL list

5.51 The quality of the lists will also be dependent on the accurate inclusion of CSAM URLs. Service providers can deploy tools to match URLs in their search content with known CSAM URLs included in a list. This measure includes elements designed to ensure that CSAM URLs are accurately included in the list, which will substantially mitigate the risk of content being incorrectly identified as a CSAM URL (see paragraph 5.73 and 5.74). We detail this risk in the 'Costs and risks' section.

5.52 In addition to the accuracy of URLs, the measure's effectiveness may be impacted by the type of content captured in the URL list and subsequently detected and removed from search results. Responding to the November 2023 Consultation, some stakeholders suggested the measure would not effectively address some kinds of harm, noting that it does not capture all content types relevant to CSAM (such as livestreaming footage or [✂]).[865] This feedback is outlined in paragraph 5.15. We recognise that this measure does not capture livestreams or [✂] which, as a result, could limit the effectiveness of this measure to identify and remove CSAM.

- **Livestreaming:** We recognise that livestreaming functionalities could be exploited either for streaming of on-demand contact abuse or recorded CSAM. However, our current understanding is that using URL lists is unlikely to be an effective mechanism to address this harm. Most livestreams are ephemeral and will have ended before they can be added to URL lists. In addition, perpetrators are likely to obtain access to livestreams through bilateral messaging, trusted groups, or advertising on social media sites (rather than via search services). However, any websites dedicated to hosting CSAM livestreams could be added to a URL list and removed from search results in accordance with this measure.
- [✂].[866] [867]

### Effective maintenance of the URL lists

5.53 As well as ensuring that URLs are accurately identified as CSAM URLs, it is important that the lists are effectively maintained. The measure sets out that service providers should ensure that arrangements are in place to regularly update the list with identified CSAM URLs, and regularly review listed CSAM URLs and remove any which are no longer CSAM URLs, as well as to secure the list from unauthorised access, interference or exploitation. It also recommends the service provider to regularly monitor the list for the purpose of removing search content relating to listed CSAM URLs. These elements of the measure are designed to ensure its effective implementation, allowing for webpages identified as hosting CSAM to be removed from search services' search content in a timely way.

## Benefits and effectiveness conclusion

5.54 Overall, we consider that this measure provides significant benefits by reducing the harm caused by the risk of exposure to CSAM (for a detailed explanation of these risks, refer to the Register chapter titled 'CSEA'). We expect this measure will significantly reduce access to CSAM via search services and as a result, consider that this measure will be an effective means of addressing the harm that we have identified.

[865] [✂].

[866] [✂].

[867] [✂].

5.55 As discussed in paragraph 5.32, we maintain that the measure’s flexible approach to automated moderation of search results offers the most beneficial option for providers. We consider that this approach:

- avoids ambiguity and possible confusion that could result from service providers defining the relevant terminology differently;
- provides flexibility for the implementation and application of the measure due to its lack of prescriptiveness regarding automated systems and processes;
- creates opportunities for the measure to evolve with technical advancement as service providers develop new automated moderation systems and processes; and
- correlates with the scope of the Act by prescribing outcomes for UK users (while allowing providers to take a wider approach, such as removing content from global search indexes).

## Costs and risks

### Costs

5.56 The implementation of this measure is expected to give rise to costs for service providers ranging from sourcing and integration of the list to software development and maintenance.

5.57 In our November 2023 Consultation, we identified that the main costs for service providers would be in relation to obtaining an appropriate URL list, ensuring their system acts on the list to remove matched URLs from their search results, and ensuring that URLs no longer included in a list are reinstated. There are likely to be several specific costs relating to these steps.

5.58 Search service providers will be expected to source a URL list from at least one third party, and it is likely that a cost would be associated with this, such as a fee to support the work of maintaining the list. For example, the IWF charges providers an annual fee ranging from over £1,000 to over £90,000 per year depending on industry sector and company size.[868]

5.59 Service providers are expected to integrate those list(s) of URLs into their existing systems and regularly test their index against the latest version of the list. This approach mirrors the steps some providers have already undertaken to remove content that infringes on copyright.[869] [870]

5.60 The system costs are expected to include both the initial software development cost and an ongoing cost of maintaining the technology. Areas of software development include authentication, identity lifecycle management, storage, user interface, workflow, messaging, testing, and security.

5.61 In response to our November 2023 Consultation, we did not receive any concerns from stakeholders on the main costs identified with this measure or our estimation of these costs.[871] While we have made some changes based on stakeholder feedback we received

[868] Internet Watch Foundation. Membership Fees. [accessed 16 October 2024].

[869] Intellectual Property Office, 2017. Search Engines and creative industries Sign anti-piracy agreement. [accessed 16 October 2024].

[870] UK Government, 2017. Voluntary Code of Practice on Search and Copyright. [accessed 16 October 2024].

[871] We did however receive some feedback on our general cost assumptions (e.g. salary assumptions) that fed into these costs. We consider that feedback in Annex 5.

regarding other aspects of the measure (as mentioned in paragraphs 5.24), these are relatively minor amendments to strengthen or clarify existing elements of the measure and we do not expect them to lead to significant additional costs.

5.62 We estimate that implementing this type of functionality would take approximately two to 16 months of software engineering time. We have also included equivalent time for other professional occupation staff. Based on the labour cost assumptions set out in Annex 5, we expect the initial implementation cost would be somewhere between £20,000 and £300,000, depending on the complexity and size of the service's existing system infrastructure.[872]

5.63 We expect the ongoing costs would include:

- ongoing access to the search engine infrastructure;
- additional recurring software licensing costs;
- costs of installing new infrastructure;
- any recurring annual fee to a third party to access their URL list; and
- the costs of securing the URL list from unauthorised access.[873]

5.64 Consistent with our standard assumption for the ongoing costs of system changes, we assume the annual running costs are 25% of the original implementation costs.[874] We therefore estimate the annual running costs to be approximately £5,000 to £80,000 per annum. This is in addition to any annual fee to the third party providing the URL list.

5.65 The overall costs of implementing and applying this measure are likely to vary as system infrastructure differs considerably from service to service. We expect service providers will predominately use automated systems to implement this measure (given the frequency of additions and removals of URLs within the lists).[875] We understand that two providers of large search services (Google and Microsoft) already work with established organisations to source URL lists and remove CSAM from their indexes. We therefore do not expect these providers to incur additional upfront costs to apply this measure.

5.66 We considered the impact of our measure on entities involved in downstream search service arrangements.[876] Based on current practices, [✂].[877]

5.67 We therefore expect there to be zero or negligible additional costs associated with implementing this measure for the relevant entities involved in a downstream search service arrangement, where the downstream search service obtains search results from Microsoft Bing or Google Search's indexing operations. We discuss our position on who the

[872] This is the same as in the November 2023 Consultation, except that we have updated the estimates in line with the latest wage data released by ONS. Since our cost estimates are rounded, the resulting estimates may not necessarily change.

[873] We recognise that some of these costs may be higher for service providers which access and ingest CSAM URL lists from more than one third party.

[874] See Annex 5 for an explanation of this assumption.

[875] To the extent that human involvement exists, we expect this to be limited to decisions as to whether to reinstate search content. The actual removal of results relating to URLs or content associated with those URLs would be an automated process.

[876] Competition and Markets Authority, 2020, "Online platforms and digital advertising - Market study final report" 1 July 2020; paragraphs 3.6-3.7 and 3.80-3.86.

[877] [✂].

'provider' of a downstream general search service is in the 'Our approach to developing Codes measures' chapter.

5.68 We recognise, however, that there are some smaller general search services that carry out their own indexing[878], and which may incur the costs set out in paragraph 5.62 to 5.64 where they do not already use CSAM URL lists. We are aware of at least one small general search service (Mojeek) that carries out its own indexing and uses CSAM URL lists. We do not have information on whether any other such providers currently use CSAM URL lists, such as small general search services that focus on non-English speaking users (for example, Yandex and Baidu) that may fall within scope of this measure.[879]

5.69 We also acknowledge the concern that this measure could create additional barriers to service providers (particularly smaller providers) entering the market that plan to conduct their own indexing, as these providers would incur the costs associated with this measure (which may represent a significant barrier to entry). However, as mentioned above, we are aware of at least one smaller provider (Mojeek) that has entered the market whilst relying on their own indexing and using a CSAM URL list, which suggests that this barrier would not necessarily prevent entry for smaller and new platforms.

5.70 Any search service operating in Australia that is subject to the eSafety Search Code would be required to "take appropriate steps to delist or block search results that contain known CSAM". Where relevant search services are using CSAM URL lists to meet this requirement, they may only be subject to minimal costs (if any at all) when implementing this measure.[880]

### Risks

5.71 We recognise there are a number of risks associated with this measure, including:

- incorrect identification and inclusion of CSAM URLs and subsequent unjustified removal of search content;
- unauthorised access and/or changes to the URL list; and
- continued removal of search content despite the URL no longer containing or depicting CSAM.

#### Risk of incorrect identification, inclusion, and removal of CSAM URLs

5.72 As this measure involves the removal of search content, there is a risk of content being wrongly removed from services as a result of a URL being incorrectly identified as a CSAM URL. We consider this risk to be limited because the measure recommends service providers to source one or more lists of CSAM URLs from organisations with expertise in the identification of CSAM, and our expectation is that these lists will be highly accurate. The IWF have a complaints procedure to allow individuals to appeal the inclusion of a URL on their list. For example, in 2020, the IWF reported that it received only 11 complaints about the incorrect inclusion of URLs in their CSAM URL list, and only one of these was upheld.[881]

[878] Competition and Markets Authority, 2020, paragraph 3.54 footnote 104.

[879] We note that these services may be small in the context of the UK. The application of this measure would depend on whether they fall in scope of the regime.

[880] eSafety, Internet Search Engine Services Online Safety Code (Class 1A and Class 1B Material), paragraph 7(2)(a).

[881] Internet Watch Foundation. Complaints. [accessed 16 October 2024].

In 2021, there were four of these complaints, and none were upheld.[882] [883] Furthermore, as outlined in paragraph 5.44, there are reputable lists that are regularly updated to ensure the accuracy of content. For example, the IWF list is updated twice a day, including removing and adding URLs. The regular maintenance of lists, as expected by the measure, ensures accurate content will be captured.

5.73 The design of this measure seeks to maximise the accuracy of URL lists to reduce the risk of inaccurate removal of content. It sets out requirements that should be met by the organisation providing the list for it to be appropriate to use. The service provider should assure itself of these when sourcing a list. These require arrangements to:

- ensure that suspected CSAM URLs are correctly identified before they are added to the list;
- regularly update the list with identified CSAM URLs;
- regularly review listed CSAM URLs and remove from the list any which are no longer CSAM URLs (with the aim of ensuring that URLs do not remain on the list for longer than is needed); and
- secure the list from unauthorised access, interference, or exploitation (which will ensure the lists are not altered by unauthorised individuals to wrongly include URLs or remove URLs containing CSAM).

### Risk of unauthorised access and/or changes to the URL list

5.74 We recognise that the URL list itself may be vulnerable to security compromises, including unauthorised access, interference and exploitation. This risk was reiterated by stakeholders in responses to the November 2023 Consultation as outlined in paragraphs 5.17 and 5.18. Perpetrators may attempt to attack services to make measures less effective and the simpler the implementation of the technology, the higher the risk that the service may be attacked by perpetrators to gain access to the URL list in question. Examples of potential security risks may include [✂].[884] Having considered stakeholder responses commenting on the need to secure the list, we have made a minor adjustment to the measure to strengthen general security standards by requiring service providers to put in place a policy for its security measures to mitigate the risks of unauthorised access, interference and exploitation.

### Potential biases in hash databases

5.75 We also recognise there is a risk of bias in relation to the compilation of the URL lists[885], which could disproportionately impact women and girls.[886] The risk of bias was noted by two stakeholder, as outlined in paragraph 5.19.[887] This concern was also raised in relation to our proposals on automated content moderation for U2U service providers.[888] We acknowledge a risk of biases could arise when identifying suspected CSAM URLs, including in relation to which websites are scrutinised or how suspected CSAM URLs are surfaced (for

[882] Internet Watch Foundation, IWF Annual Report 2021 [accessed 06 November 2024.]
[883] No equivalent data from 2022 and 2023 was located.
[884] [✂].
[885] OSTIA response to November 2023 Consultation, p.14.
[886] Glitch response to November 2023 Consultation, p.8.
[887] Glitch response to November 2023 Consultation, p.8.
[888] Glitch response to November 2023 Consultation, p.8; OSTIA response to November 2023 Consultation, p.14.

example, by the use of automated methods such as hash matching, tools based on artificial intelligence or machine learning, proactive searching by human analysts, or reactively in response to user reports). A risk of biases could also arise regarding how the suspected URLs are subsequently assessed as CSAM (for example, by age determination through human assessment). This may create biases that underrepresent the scale and nature of the problem of CSAM for different ages and minority groups. We consider that these risks are likely to be mitigated by the elements which promote the accuracy and effectiveness of the list (see paragraphs 5.29-5.31). However, we have responded to these concerns by revising the measure to add that the service provider should ensure that the arrangements for identifying or assessing suspected CSAM URLs do not plainly discriminate on the basis of protected characteristics (within the meaning of Part 2 of the Equality Act 2010), such as sex or race.[889] [890]

### Risk of continued removal of search content

5.76 There is also a risk of URLs continuing not to appear in search results after CSAM is removed (or where the URL was wrongly included in the list in the first place and subsequently removed from the list). As well as impacting users' ability to access websites through search results, this could have significant commercial impacts on website owners.

5.77 The design of the measure seeks to mitigate this risk by (1) ensuring the organisation providing the list has arrangements in place to regularly review listed CSAM URLs and remove any from the list when no longer CSAM URLs, and (2) ensuring the service provider regularly monitors the list and ensures that action taken in relation to a URL is swiftly reversed once it is removed from the list (unless the provider considers this would be inappropriate). In addition, "interested persons"[891] can complain to service providers where content relating to them no longer appears in search results as a result of this measure. Service providers should take appropriate action in response to such a complaint. Such action could include asking the organisation providing the list to review the inclusion on the list of the website in question. We have included the relevant Codes measures about appeals by interested persons as safeguards for users' and interested persons' freedom of expression.

5.78 We also recognise the value of notifying website operators where a URL on their website has been added to the list, which was raised by Glitch in response to the November 2023 Consultation.[892] We outline their response in paragraph 5.20. We understand that website owners or hosting providers will usually be notified in such cases, given that the providers of URL lists are typically child protection organisations or law enforcement authorities which aim to ensure CSAM is taken down. Where the URL is hosted outside the UK, this will usually be done by an organisation in the relevant country.[893]

[889] Similar changes have been made to the measures set out in chapter 4 of this Volume: 'Automated content moderation'.

[890] To note, child protection organisations are also alive to the risk of biases. However, we recognise that a risk of biases in the URL lists sourced by service providers remains.

[891] See footnote 903.

[892] Glitch response to November 2023 Consultation, p.9.

[893] The IWF states that "Notifying the website owner or hosting provider of any blocked URL is the responsibility of the hotline or relevant law enforcement agency in the country believed to be hosting the content". Source: IWF. URL Blocking FAQs. [accessed 25 September 2024].

5.79 Overall, we consider that the way we have designed this measure includes suitable mitigations for the risks we have identified and that these risks can be managed. As discussed in this section, risk levels can be reduced and mitigated through the practices of both service providers implementing the measure and organisations which compile and maintain URL lists. We have therefore designed the measure with a number of safeguards detailed in this chapter and adjusted some of the safeguards following our stakeholder feedback to our consultation. We assess the measure's impacts on rights in the next section.

## Rights impact

5.80 This section considers the measure's impacts on rights under Articles 8 and 10 of the European Convention on Human Rights ('ECHR').

5.81 As explained in 'Introduction, our duties, and navigating the Statement', Article 10 of the ECHR sets out the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 8 of the ECHR sets out the right to respect for individuals' private and family life.

5.82 In essence, restrictions on those rights must be necessary and proportionate – that is, the measure's contribution to its objective must outweigh its adverse impacts.

5.83 Our assessment of the adverse impacts of the measure is therefore to be balanced against the measure's contribution to its objective of reducing harm associated with accessing CSAM in or via search results. Parliament has legislated for CSAM to be designated as "priority illegal content" under the Act, requiring search service providers to implement systems and processes designed to minimise the risk of individuals encountering such content in search results. This reflects the very substantial public interest in measures that reduce access to CSAM online, relating to each of the prevention of crime, the protection of health and morals, and the protection of the rights of others. The 'Benefits and effectiveness' section discusses the measure's effectiveness and the associated benefits in more detail.

### Freedom of expression

5.84 An interference with the right to freedom of expression must be prescribed by law, pursue a legitimate aim, be proportionate to that legitimate aim and correspond to a pressing social need.

5.85 We recognise that the removal of URLs from search results has the potential to constitute a significant interference with the rights of website/database providers to impart information and the rights of users to receive it. It can also interfere (though to a lesser degree) with the rights of search service providers to impart information. The removal of results relating to a URL from search results means, in practice, that users will no longer be able to encounter it on or via that service. This would affect any legal content hosted at that URL, as well as illegal content. Given the importance of search services to internet users' navigation of the internet, this would have a significant effect on users and website providers even though affected webpages would remain directly accessible. Whether there is a sufficient level of certainty as to the illegal nature of the content is therefore important in considering if the interference is justified.

5.86 CSAM is an extremely harmful kind of illegal content. Where CSAM URLs are correctly identified and cannot be encountered in or via search results, the content contained at that

URL either does not engage Article 10 ECHR or the restrictions in relation to that content are clearly justified to protect overriding public interests – namely the prevention of crime, protection of morals, and protections of rights of others (in particular, the rights of the children concerned).

5.87 We expect there to be few cases where search results for URLs that do not include CSAM are removed. Our expectation is that the automated systems and processes used to identify search content that relates to listed URLs will be highly accurate. As explained in paragraph 5.72 and 5.73, whether removed search content is a CSAM URL will therefore depend on the accuracy of the URL list.

5.88 The design of the measure includes a number of provisions to ensure that CSAM URLs are accurately included on the list. In turn, these operate as safeguards to protect users' and website/database owners' freedom of expression. The measure sets out the need for service providers to ensure that the person (or persons) from whom it has sourced the list (or lists) has arrangements in place to:

- ensure that suspected CSAM URLs are correctly identified before being added to the list;
- regularly review URLs on the list and remove any which are no longer CSAM URLs; and
- secure the list from unauthorised access, interference or exploitation.

5.89 The measure also sets out that service providers should have an appropriate policy in place to secure any copy of a list (or lists) held for the purposes of the measure from unauthorised access, interference, or exploitation, and that measures are taken in accordance with that policy.

5.90 In cases where a URL or domain is removed from a CSAM URL list, service providers should swiftly reverse the action taken in relation to that URL or domain (unless they consider that it would be inappropriate to do so). This provides a further safeguard for freedom of expression.

5.91 However, we acknowledge that there may still be a small number of cases where a URL has been incorrectly included on a list as a CSAM URL, as well as cases where a URL continues to be removed from search results for a period after the CSAM has been removed.

5.92 In such cases, complaints procedures operated in accordance with section 32 of the Act should allow interested persons to complain and for appropriate action to be taken in response (see chapter 6 of this Volume: 'Reporting and complaints'). The measure specifies other Codes measures as safeguards for freedom of expression, including:

- Enabling interested persons to complain if measures taken in accordance with the illegal content safety duties result in their content no longer appearing in search results.
- Ensuring that providers' publicly available statements give information about the proactive technology used and the policies and processes that govern the handling and resolution of complaints.

5.93 These Codes measures help safeguard freedom of expression in a number of different ways. This includes providing transparency about the technology used and giving clear information on how to make a complaint – which, in turn, provides complainants with a mechanism for redress and a route to rectify any negative impact resulting from having their content removed from the service's search content.

5.94 We acknowledge that there may be some interference with freedom of expression if the content present at a URL includes legitimate content as well as CSAM. The measure recommends that URLs should be listed where the relevant domain is entirely or predominantly dedicated to CSAM. We recognise that this could have some impact on users' rights to freedom of expression, as well as on the rights of website/database owners who make such content available and the rights of service providers who make such information available to their users. However, we consider that this is justified to protect public interests due to the risk that users accessing such URLs will go on to encounter CSAM.

### Privacy

5.95 An interference with the right to privacy must be in accordance with the law, pursue a legitimate aim, be proportionate to the legitimate aim and correspond to a pressing social need.

5.96 We consider the measure's impact on users' rights to privacy under Article 8 ECHR to be minimal. The measure is directed at removing search content from the content that can be encountered in or via search results. It does not involve analysis of users' search requests.

5.97 The impact on the privacy of those responsible for websites or databases that are included in search content is also limited. These websites or databases are already publicly available on the internet and, in any event, the measure does not involve the search service provider accessing or analysing the content or recording any information about the person responsible for it. Instead, it relies on the assessment carried out by the organisation which maintains the URL list.

5.98 Where CSAM is detected, service providers may in certain circumstances be required (or choose) to report this to a law enforcement authority (or to a designated reporting body such as National Center for Missing & Exploited Children). Relevantly, section 66 of the Act makes provision which, when brought into force, will require providers of regulated search services to report to the NCA detected CSEA content present on websites or databases which is not otherwise reported.[894] However, unlike moderation processes which may result in the service provider itself detecting CSEA content present at URLs indexed by a service, the measure involves the use of URL lists sourced from organisations (such as child protection organisations) which are likely to already engage with law enforcement authorities when they identify CSAM.

5.99 The measure involves the assessment of content at suspected CSAM URLs by organisations providing URL lists, which will involve the processing of personal data (as in, information which relates to an identified or identifiable person) of victims and survivors and others. Such work is already undertaken by various child protection organisations and law enforcement authorities, but our measure could result in additional processing taking place. We expect these organisations to have robust security and to ensure that any processing is carried out in accordance with data protection law, which will safeguard against privacy risks. Overall, victim and survivor rights will be safeguarded by the measure because it will help reduce access to CSAM depicting them.

[894] In the case of a non-UK provider of a regulated search service, the duty is limited to "UK-linked" CSEA content: s.66(4).

5.100 Overall, we consider that the impact of the measure on rights under Articles 8 and 10 of the ECHR are proportionate to the measure's aim of reducing users' exposure to CSAM in search results.

### Who this measure applies to

5.101 In the November 2023 Consultation we set out our provisional view that the measure we had proposed should apply to all general search services, irrespective of size.

5.102 Following the consultation, we have decided to proceed with the approach we proposed. We conclude that our approach is proportionate considering the scale and severity of CSAM online, our analysis of the effectiveness of the measure, the costs to service providers of implementing it, and its impact on users' and website/database providers' rights.

5.103 This scope of general search service providers covers two types of services:

- large general search services (such as Google Search and Microsoft Bing); and
- smaller general search services, including those that constitute a type of search service called a downstream service (such as Yahoo or Ecosia).[895]

5.104 We consider that any general search service carries a risk that users will encounter CSAM in or via search results. An important consideration in our decision to apply the measure to all general search services was the risk of perpetrators otherwise exploiting service providers not in scope of the measure as a means to circumvent safety measures implemented on other search services. We do not expect the costs of implementing and maintaining this measure to be significant for most search services, including for smaller search services [see paragraphs 5.61 to 5.65]. Therefore, we consider it proportionate to apply the measure across all general search services.

5.105 In the November 2023 Consultation, we explained that there was a lack of evidence to suggest that vertical search services were used to disseminate any type of priority illegal content (in this case, CSAM).[896] As mentioned in paragraph 5.22, one stakeholder did question why vertical search services were excluded from the proposed measure.[897] [898] However, we have not seen clear evidence (submitted in consultation responses or otherwise) to suggest that vertical search services play a significant role in the dissemination of CSAM. We therefore remain of the view that it is appropriate to apply the measure only to general search services.

## Conclusion

5.106 We consider this measure to be an effective means to reduce the risk of users of search services encountering CSAM in search content of the service, with substantial benefits.

5.107 Having considered the costs, risks and associated impacts on the rights of users, website/database owners, and providers of search services, we consider it to be a

[895] We discuss who the 'provider' of a downstream general search service is in our chapter 'Our approach to developing Codes measures'.

[896] Paragraph 15.84 of Volume 4 of the November 2023 Consultation, p.167.

[897] WeProtect Global Alliance response to November 2023 Consultation, p.17.

[898] In the November 2023 Consultation, we explained that vertical search services (as we define them) do not use a search index – as such, a measure that recommended "deindexing" of CSAM URLs would not be applicable to them. As the measure no longer refers to "deindexing" but rather to removing results related to listed CSAM URLs, this point no longer applies.

proportionate safety measure to recommend providers of general search services take for the purpose of compliance with their illegal content safety duties (in particular, the duty under section 27(3)(a) to operate the service using proportionate systems and processes designed to minimise the risk of individuals encountering search content that is priority illegal content).

5.108 We have also designed the measure to incorporate a number of safeguards for the protection of the right of users and interested persons to freedom of expression.

5.109 We have therefore decided to include this measure in the Illegal Content Codes of Practice for search services. It is referred to as ICS C7 and is part of the CSEA Code of Practice.

# 6. Reporting and complaints

## What is this chapter about?

This chapter sets out the measures we are recommending in relation to reporting and complaints, why we are recommending them, and to which user-to-user (U2U) and search services they should apply.

## What decisions have we made?

We are recommending the following measures:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| ICU D1 / ICS D1 | All providers of U2U and search services should have **complaints systems and processes** that enable prospective complainants (UK users, affected persons and (for search services) interested persons) to make relevant complaints in a way that will secure appropriate action in relation to them. | Providers of all services. |
| ICU D2 / ICS D2 | Providers should design and operate complaints procedures so that they are **easy to find, access, and use**. | Providers of all services. |
| ICU D3 | Providers should ensure their reporting tool for specific content enables to access information that **informs complainants** if they will share information about a complaint **with another user**, and what information will be shared. This includes information relating to the original complaint and complainant if the other user subsequently appeals. | • Providers of large U2U services which are likely to be accessed by children.<br>• Providers of U2U services at medium or high risk of any kind of illegal harm, which are likely to be accessed by children. |
| ICU D4 / ICS D3 | Providers should **acknowledge receipt of a complaint** and provide complainants with an **indicative timeframe** for when a complaint might be resolved. | • Providers of large U2U services.<br>• Providers of large general search services.<br>• Providers of services at medium or high risk of any kind of illegal harm. |

| | | |
|---|---|---|
| **ICU D5 / ICS D4** | Providers should inform a complainant about the **possible outcomes of a complaint**, including whether the service will update the complainant on the outcome. | • Providers of large U2U services likely to be accessed by children.<br>• Providers of large general search services likely to be accessed by children.<br>• Providers of services at medium or high risk of any kind of illegal harm, which are likely to be accessed by children. |
| **ICU D6 / ICS D5** | Providers should give complainants the **option to opt out of receiving non-ephemeral communications** about a complaint from the service provider. | • Providers of large U2U services.<br>• Providers of large general search services.<br>• Providers of services at medium or high risk of any kind of illegal harm. |
| **ICU D7 / ICS D6** | Providers should handle complaints about suspected illegal content in accordance with their **content prioritisation processes** and content moderation functions, **or promptly** if the recommendations about prioritisation and targets do not apply to that provider. | Providers of all services. |
| **ICU D8 / ICS D7** | Providers should determine **complaints which are appeals** and monitor against performance targets for time taken to determine and for accuracy. The provider should have a prioritisation policy for appeals. | • Providers of large U2U services.<br>• Providers of large general search services.<br>• Providers of multi-risk services. |
| **ICU D9 / ICS D8** | Providers should determine **complaints which are appeals** promptly. | • Providers of U2U services that are neither large nor multi-risk.<br>• Providers of search services that are neither large general search services nor multi-risk. |

|  |  |  |
|---|---|---|
| ICU D10 / ICS D9 | For **complaints which are appeals**, if a provider reverses a decision that content was illegal content, it should: **reverse the action** taken (so far as appropriate and possible); **adjust any relevant moderation guidance** if appropriate; and take appropriate steps to secure that the use of automated moderation technology does not result in the **same content being taken down** / search content no longer appearing in search results or being given a lower priority in the ranking of search results, again. | Providers of all services. |
| ICU D11 | For complaints about the use of proactive technology in breach of terms of service, providers should **inform complainants of the action the provider might take and the complainant's right to bring proceedings.** | Providers of U2U services. |
| ICS D10 | For complaints about the use of proactive technology in breach of policies, providers should **inform complainants of the action the provider might take**. | Providers of search services. |
| ICU D12 / ICS D11 | Providers should **nominate an individual or team** to ensure that **all other relevant complaints** are directed to the appropriate individual or team for processing. | Providers of all services. |
| ICU D13 / ICS D12 | A provider may disregard a relevant complaint (that is not an appeal) if it has a policy that sets out the information and attributes that would indicate that a relevant complaint is manifestly unfounded. It must make decisions in accordance with this policy and review the application of the policy annually. | Providers of all services. |
| ICU D14 / ICS D13 | Providers should establish and maintain a **dedicated reporting channel for trusted flaggers** (including at least the specified list of public bodies) to use to report, at a minimum, fraud. | • Providers of large U2U services that are at medium or high risk of fraud.<br>• Providers of large general search services that are at medium or high risk of fraud. |

### Why have we made these decisions?

Complaints are important mechanisms for providers to become aware of harmful content. The decisions we have taken today will help ensure that reporting and complaints functions operate effectively. We consider this will make providers better able to identify and remove illegal content, thereby reducing harms to users. We have included a provision in our Codes allowing providers to disregard complaints which are manifestly unfounded ('spam' complaints). This exception means providers can focus their resources on taking appropriate action against legitimate complaints, and do not need to review complaints that are part of a co-ordinated attack or have been submitted by malicious actors.

Dedicated reporting channels provide an easy way for expert 'trusted flaggers' to report problems to providers. These can play a valuable role in improving detection of illegal content, therefore reducing harm to users. In principle dedicated reporting channels could be used to address a wide range of harms. In this first version of our Codes we have focused our recommendations regarding dedicated reporting channels for trusted flaggers on fraud. That is because we have received specific evidence indicating that organisations with expertise in fraud often find it difficult to report known scams to providers and that the creation of a dedicated reporting channel would play an important role in addressing this problem.

# Introduction

## Duties under the Act

6.1 Reporting and complaints processes help service providers take action on harmful or illegal content. The Online Safety Act 2023 ('the Act') requires service providers to process different complaints depending on whether the service is a user-to-user ('U2U') or a search service.[899] They are referred to collectively as 'relevant complaints'. Relevant complaints include:

- complaints by UK users and affected persons about content they consider to be illegal, [900] [901]
- complaints by UK users and affected persons who believe the provider is not complying with duties set out in the Act,[902]
- complaints by a UK user who is appealing a decision taken by the provider, on the basis of suspected illegal content, that has resulted in their use of a U2U service being restricted or their content being taken down, or (on a search service) complaints by an

[899] It includes duties for providers of U2U services in sections 20 and 21 and for providers of search services in sections 31 and 32.

[900] Sections 21(4)(a) and 32(4)(a); see also sections 20(2) and 31(2)) of the Act. 'User' is interpreted in accordance with sections 8(3)(b) and 25(1)(c) of the Act.

[901] The Online Safety Act (2023) defines 'affected person' as a 'person, other than a user of the service in question, who is in the United Kingdom and who is (a) the subject of the content, (b) a member of a class or group or people with a certain characteristic targeted by the content, (c) a parent of, or other adult with responsibility for, a child who is a user of the service or is the subject of the content, or (d) an adult providing assistance in using the service to another adult who requires such assistance, where that other adult is a user of the service or is the subject of the content,' (sections 20(5) and 31(5) of the Act).

[902] Sections 21(4)(b) and 32(4)(b) of the Act.

interested person[903] who is appealing a decision taken by the provider, on the basis of suspected illegal content, that has resulted in their content being removed or given a lower priority in search results,[904]

- complaints by a UK user or interested person regarding the use of proactive technology that a user considers to be in breach of a provider's terms of service or policies on use, which impacts the prominence of a user's/interested person's content.[905]

6.2 The Act has specific provisions for providers of U2U and search services about establishing and operating complaints procedures. Specifically:

6.3 In accordance with these duties, the measures in this chapter concern:

- the design of complaints procedures;
- the prioritisation and handling of all relevant complaints; and
- the appropriate action a provider should take in response to relevant complaints, including appeals.

6.4 The measures discussed in this chapter will assist service providers in operating a robust and user-friendly complaints procedure, ensuring that service providers take appropriate action in relation to a complaint.

6.5 We encourage service providers to regularly and systematically incorporate the findings from their risk assessments when considering how best to act in accordance with and implement our measures. This will help ensure that complaints procedures are designed to meet the specific requirements of a service's UK userbase and the kinds of illegal harm that a service provider has assessed on its service.

6.6 Duties regarding relevant complaints about content that is harmful to children will be addressed in the Protection of Children Statement.[906]

## Chapter structure

6.7 In our November 2023 Illegal Harms Consultation ('November 2023 Consultation'), we proposed a package of measures across U2U and search services. We updated some of these measures in our May 2024 Consultation on Protecting Children from Harms Online ('May 2024 Consultation'), and proposed two further measures. The package of measures was:

a) All providers of U2U and search services should have complaints processes for UK users, affected persons and (for search services) interested persons to make relevant complaints in a way that will secure the provider will take appropriate action in relation to them.
b) All providers of U2U and search services should operate a complaints system for all types of relevant complaints that is easy to find, access, and use.

[903] The Act defines an interested person in relation to a search service as "a person that is responsible for a website or database capable of being searched by the search engine, provided that (a) in the case of an individual, the individual is in the United Kingdom; (b) in the case of an entity, the entity is incorporated or formed under the law of any part of the United Kingdom", (section 227(7) of the Act).
[904] Sections 21(4)(c), 21(4)(d), and 32(4)(c) of the Act.
[905] Sections 21(4)(e) and 32(4)(d) of the Act.
[906] Section 21(5) and Section 32(5) of the Act.

c) Providers of U2U services likely to be accessed by children should ensure that users and affected persons can easily access information on whether the provider will share information about a complaint with another user, and what information will be shared.

d) All providers of U2U and search services should acknowledge receipt of a complaint and provide the user with an indicative timeframe for when the complaint might be decided.

e) Providers of U2U and search services likely to be accessed by children should inform a complainant about the possible outcomes of a complaint, including whether the service will update the complainant on the outcome of the complaint, in their acknowledgement of a complaint.

f) All providers of U2U and search services should handle relevant complaints about suspected illegal content in accordance with our proposed moderation.

g) recommendations (including where relevant as to prioritisation and performance targets) (or promptly if these do not exist).

h) Providers of U2U services which are large or multi-risk, and of search services which are large general search services or multi-risk should determine relevant complaints which are appeals against performance targets). When determining what priority to give to its review of an appeal, the provider should consider: the seriousness of the action taken against the user or content; whether the initial decision was made by content identification technology; and the past error rate for illegal content judgements.

i) Providers which are neither large nor multi-risk should determine appeals promptly.

j) All U2U and search service providers should, if they reverse a decision that content was illegal content, reverse the action taken against the user or the content (so far as appropriate). It should also adjust the relevant moderation guidance where necessary to avoid similar errors in future, and where applicable and necessary take steps within its power to secure that the use of automated content moderation technology does not cause the same content to be taken down again.

k) All providers of U2U and search services should inform complainants of the complainant's rights to bring proceedings (where relevant) if they believe that the use of proactive technology has resulted in: content being taken down, given a lower priority or access to it being restricted; search content being deindexed or downranked; or if the technology has been used in a way that is in breach of the service provider's terms of service or publicly available statement.

l) All providers of U2U and search services should establish a triage process for relevant complaints, which should be dealt with by the most relevant function or team and in a way that protects users and the provider's compliance with other applicable laws, and within appropriate timeframes.

m) Providers of large U2U services with a high or medium risk of fraud, and large general search services with a medium or high risk of fraud, should establish and maintain a dedicated reporting channel for specified public bodies (trusted flaggers) to use.[907]

6.8 In this chapter, we set out and explain our decisions to include measures in the Illegal Content Codes of Practice for U2U and search services relating to reporting and complaints.

[907] In our November 2023 Illegal Harms Consultation, there was a discrepancy between the draft Code and the consultation chapter on the application of this measure for search services. Our reasoning in our consultation document talked about "large services" rather than, as per the proposed Code, "large general search services". Our approach throughout our November 2023 Consultation was to propose measures for large general search services rather than for all large search services as we were not aware of evidence of a risk of fraud or any other harm on vertical search services.

We have made some changes or clarificatory amendments to all the reporting and complaints measures we proposed and have also included some new provisions in response to stakeholder feedback on our measures.

6.9 This chapter generally discusses one measure per section, and we largely work through the measures in the order set out above, with the following exceptions:

- The measure relating to easy-to-use, easy-to-access, and accessible complaints procedures comprises several components that speak to different duties and objectives.[908] Because of this, we consider the measure across two separate sections.
- We consider measures relating to appropriate action in response to relevant complaints in one section because we received cross-cutting feedback in response to these measures.[909]
- Similarly, we consider measures relating to complaints which are appeals in one section because we received cross-cutting feedback in response to these measures.[910]

# Measure on enabling complaints

6.10 In our November 2023 Consultation, we recommended that all providers of U2U and search services should have complaints processes for UK users, affected persons and (for search services) interested persons to make relevant complaints in a way that ensures the provider will take appropriate action in relation to them.[911]

6.11 We stated that service providers would need to have a minimum of two reporting tools: a standard reporting tool for making complaints about content, and a tool to submit other relevant complaints such as complaints about the efficacy or operation of a reporting function. Other types of complaints could be made using one of these tools, or other tools.[912]

6.12 This measure is the minimum action necessary to comply with the requirement in the Act to operate a complaints procedure.

## Summary of stakeholder feedback[913]

6.13 Stakeholders were generally supportive of this measure.[914]

[908] In our Codes, the relevant measures are ICU D2 for U2U services and ICS D2 for search services.

[909] In our Codes, the relevant measures are ICU D7, ICU D11 and ICU D12 for U2U services. The equivalent search measures are ICS D6, ICS D10 and ICS D11, respectively.

[910] In our Codes, the relevant measures are ICU D8-D10 for U2U services and ICS D7-D9 for search services.

[911] For the definition of "interested persons" see section 227(7) of the Act.

[912] The final measure discussed in this chapter requires large U2U and large general search services at high or medium risk of fraud to establish a reporting channel for trusted flaggers to submit reports about fraud. This channel is in addition to the minimum of two reporting tools we are recommending as part of this measure for relevant services.

[913] Note: this list is not exhaustive and further responses can be found in Annex 1.

[914] 5Rights Foundation response to November 2023 Illegal Harms Consultation, p.23; Betting and Gaming Council response to November 2023 Illegal Harms Consultation, p.9; Canadian Centre for Child Protection (C3P) response to November 2023 Illegal Harms Consultation, p.21; Cats Protection response to November 2023 Illegal Harms Consultation, p.12; Centre for Competition Policy response to November 2023 Illegal Harms Consultation, p.17; Children's Commissioner response to November 2023 Illegal Harms Consultation, p.22; Evri

6.14 We have identified several themes that emerged from stakeholder feedback:

- Submitting complaints as an ‘affected’ or ‘interested’ person or non-registered user.
- Balance between efficient complaints procedures and safety by design.
- Multiple reporting tools and/or functions.
- Privacy and data retention.

6.15 We detail comments on these themes in the following paragraphs.

### Submitting complaints as an ‘affected’ or ‘interested’ person or non-registered user

6.16 Some stakeholders highlighted that users who are not registered with or logged in to a service (and might not have access to content only available to a user who is registered or logged-in) should be able to easily report content.[915] Other stakeholders highlighted that this was especially relevant in the particular context of children who had been exposed to illegal content.[916] LinkedIn highlighted that reporting and complaints procedures for logged-out users would need to be carefully designed to prevent or minimise abuse of such features.[917] We address these concerns in paragraph 6.30-6.32 in the ‘How this measure works’ section for this measure, but note that this is cross-cutting feedback pertinent to all reporting and complaints measures.

### Balance between efficient complaints procedures and safety by design

6.17 Noting the requirement that services regulated by the Act are ‘safe by design’, several stakeholders argued that the proposed reporting and complaints measures alone were insufficient and that the burden of responsibility should not shift from service providers to users, including children.[918]

6.18 End Violence Against Women (EVAW) Coalition highlighted that reporting should not be the primary safety measure on a service as “the majority of survivors do not report” and that it

response to November 2023 Illegal Harms Consultation, p.7; Fuller, A. response to November 2023 Illegal Harms Consultation, p.18; Global Partners Digital response to November 2023 Illegal Harms Consultation, pp.16-17; [✂]; Match Group response to November 2023 Illegal Harms Consultation, p.13; Mencap response to November 2023 Illegal Harms Consultation, p.11; Meta response to November 2023 Illegal Harms Consultation, pp.27-30; Nexus response to November 2023 Illegal Harms Consultation, pp.13-14; OnlyFans response to November 2023 Illegal Harms Consultation, p.8; Open Rights Group response to November 2023 Illegal Harms Consultation, pp.3-4; Refuge response to November 2023 Illegal Harms Consultation, pp.16-17; [✂]; Snap response to November 2023 Illegal Harms Consultation, p.13; The Cyber Helpline response to November 2023 Illegal Harms Consultation, p.15; UK Interactive Entertainment (Ukie) response to November 2023 Illegal Harms Consultation, p.22; Welsh Government response to November 2023 Illegal Harms Consultation, p.4; WeProtect Global Alliance response to November 2023 Illegal Harms Consultation, pp.17-18.

[915] Global Partners Digital response to November 2023 Consultation, p.16; Meta response to November 2023 Consultation, p.28; The Cyber Helpline response to November 2023 Consultation, p.15; UK Safer Internet Centre (UKSIC) response to November 2023 Illegal Harms Consultation, pp.39-40.

[916] 5Rights Foundation response to November 2023 Consultation, p.24; Bereaved Families for Online Safety response to November 2023 Illegal Harms Consultation, p.1.

[917] LinkedIn response to November 2023 Illegal Harms Consultation, pp.13-14.

[918] 5Rights Foundation response to November 2023 Consultation, p.24; National Society for the Prevention of Cruelty to Children (NSPCC) response to November 2023 Illegal Harms Consultation, pp.14, 30-31; NWG Network (formerly The National Working Group for Sexually Exploited Children and Young People) response to November 2023 Illegal Harms Consultation, p.11.

is neither an indicator of harm nor an alternative to robust safety by design features.[919] The Institute for Strategic Dialogue similarly argued that user content reporting must complement more proactive safety by design measures.[920] eBay highlighted that its use of proactive detection technology had led to a decrease in the number of reports received through its "Report Item" functionality, highlighting the complementary nature of reporting functionalities and safe design choices.[921]

6.19 We address this feedback in paragraph 6.33 of the 'How this measure works' section.

### Multiple reporting tools and/or functions

6.20 Match Group and Meta supported our proposal that there should be complaints processes which enable users to make each of the types of relevant complaint, including complaints about any issues about the reporting function or complaints procedure itself.[922]

6.21 Four Paws UK raised concerns that some services allow reporting for specific issues only.[923] The New Zealand Classification Office recommended a "single front door" to avoid fragmentation within a reporting and complaints procedure.[924]

6.22 Mid Size Platform Group raised concerns about the number of complaints procedures we recommended, especially the disproportionate impact on small and medium platforms.[925] Skyscanner disagreed that there should at least two processes for users to submit reports and complaints, and suggested that a joint reporting and complaints function – for example, in email format – could be designed for users to make complaints about illegal content and about the complaints function itself.[926]

6.23 We discuss this feedback in the 'How this measure works' and 'Costs and risks' sections.

### Privacy and data retention

6.24 The Information Commissioner's Office (ICO) raised concerns about privacy and data retention.[927] These are addressed in paragraph 6.48-6.51 of the 'Rights impact' section.

## Our decision

6.25 We have decided to proceed with the measure broadly as proposed in our November 2023 Consultation with some clarificatory amendments. This measure now says that all service providers should have systems and processes which enable prospective complainants to make each type of relevant complaint in a way that will secure appropriate action is taken in relation to them.

6.26 The full text of the measure can be found in our Illegal Content Codes of Practice for U2U and search services and is referred to within these Codes as ICU D1 for U2U services and ICS

[919] End Violence Against Women (EVAW) Coalition response to November 2023 Illegal Harms Consultation, p.4.
[920] Institute for Strategic Dialogue response to November 2023 Illegal Harms Consultation, pp.10-11.
[921] eBay response to November 2023 Illegal Harms Consultation, p.2.
[922] Match Group response to November 2023 Consultation, p.13; Meta response to November 2023 Consultation, p.27.
[923] Four Paws UK response to November 2023 Illegal Harms Consultation, p.13.
[924] New Zealand Classification Office response to November 2023 Illegal Harms Consultation, p.8.
[925] Mid Size Platform Group response to November 2023 Illegal Harms Consultation, p.10.
[926] Skyscanner response to November 2023 Illegal Harms Consultation, pp.20-21.
[927] Information Commissioner's Office (ICO) response to November 2023 Illegal Harms Consultation, p.18.

D1 for search services. This measure is part of our Codes on terrorism, child sexual exploitation and abuse (CSEA) and other duties.

# Our reasoning

## How this measure works

6.27 Sections 21(2)(a) and 32(2)(a) of the Act require that all providers of U2U and search services offer UK users, affected persons and (for search services) interested persons a way to submit relevant complaints.[928] Relevant complaints include those about illegal content, appeals, reporting functions, non-compliance with duties, and the use of proactive technology.

6.28 Many service providers will have designed their terms of service or statements to comply with laws in the jurisdictions in which they operate or where their services are targeted. Compliance with the Act may mean that providers have different terms and statements for UK users when compared to users elsewhere in the world.

6.29 Service providers have a choice about how they comply with the Act. They may choose to meet the requirements of the Act for all their users (no matter where in the world they are located) or choose to do so only in relation to their UK users. If a service provider wishes to comply with its duties around reporting and complaints in relation to UK users only, it will need to be able to determine if a user who has submitted an illegal content complaint has been served this content in the UK.

6.30 A 'user' does not have to be registered with a service.[929] All users, regardless of whether they are registered, should be able to make various types of relevant complaints, including complaints about illegal content. As such, the duty to enable reports relates to affected or interested persons as well as to users.

6.31 We accept that affected or interested persons – or non-registered users – may not have access to all the content on a service. However, we maintain that the service provider should enable them to submit complaints. How this is done is left to the discretion of the service provider and will vary depending on the type of service in question. For example, large providers may choose to offer a complaints procedure via a web portal, such as a 'help centre' where non-registered users can submit reports or complaints. Small or low risk services may provide an email address through which users can submit reports or complaints. This will allow users who cannot view content in the same way as registered users to submit complaints about suspected illegal content.

6.32 In order to improve the readability of our Codes, we have created a new defined term – 'prospective complainants' – which means any person who is entitled to submit a relevant complaint under the Act. For U2U services, 'prospective complainants' therefore means UK users and affected persons; for search services, 'prospective complainants' means UK users, affected persons and interested persons. In our updated measures, we refer to 'prospective complainants' or 'complainants'.

6.33 As outlined in paragraph 6.17-6.19 under the 'Summary of stakeholder feedback' section, we received feedback on balancing the operation of efficient reporting procedures with safety by design features. It is our position that enabling prospective complainants to

[928] See section 8(3)(b) and 25(1)(c) of the Act for interpretation.
[929] Section 227(2) of the Act.

submit complaints does not create a burden of responsibility on users over services. Instead, it supports prospective complainants in the complaints process and better equips services to take appropriate action to address complaints. The enabling of complaints and the operation of easy-to-use complaints systems is a complement to, rather than a substitute for, steps that service providers take themselves to proactively prevent, detect and remove illegal content. The Center for Countering Digital Hate's Star Framework notes that 'accessible, effective and responsive reporting pathways' is a way of achieving safety by design, and this is the outcome these components seek to achieve.[930]

6.34 Although the Act contains separate provisions for reporting and complaints functions, a report is a type of complaint. We explained in our November 2023 Consultation that services could operate a combined reporting and complaints function where appropriate for most users and most types of complaints, so long as it is clear how to use it for each type of complaint.[931]

6.35 In our November 2023 Consultation, we said that a service provider cannot use its content reporting tool to receive all relevant complaints.[932] This is because a content reporting tool will not be able to receive complaints about, for example, the (in)effective operation of the content reporting tool itself. The service provider should offer prospective complainants at least one other means of submitting complaints aside from its standard reporting tool. This is a direct requirement of the Act.[933] This will enable the provider to receive complaints about the reporting tool and may also help alert service providers swiftly should alternative complaints systems or processes experience technical challenges. Large U2U services at medium or high risk of fraud, and large general search services at medium or high risk of fraud, have additional obligations to operate a dedicated reporting channel for trusted flaggers to use to report fraud, which should be separate from standard reporting routes.

6.36 This measure applies regardless of whether a service is end-to-end encrypted or not. We recognise that some providers operate end-to-end encrypted services and that this may affect the implementation of this measure (and subsequent measures). This measure recommends that providers of such services should also enable prospective complainants to submit complaints, and that complaint systems and processes are designed in such a way that ensures that providers take appropriate action in relation to it. In order to do this, an end-to-end encrypted service provider might, for example, automatically attach a copy of the content concerned and the content immediately around it to the complaint. We explain our position in chapter 2 of this Volume: 'Content moderation'.

## Benefits and effectiveness

6.37 We consider this measure will deliver significant benefits to prospective complainants. This measure will encourage compliance with duties set out in the Act and will contribute to the overall aim of reducing the risk of harm online.

6.38 Providing prospective complainants with routes for making reports and complaints gives providers better knowledge of the risks on their services and helps them detect illegal

[930] Centre for Countering Digital Hate, 2022. STAR Framework – CCDH's Global Standard for Regulating Social Media [accessed 10 November 2024].

[931] Ofcom, 2023. November 2023 Illegal Harms Consultation, Volume 4: How to mitigate the risk of illegal harms – the illegal content Codes of Practice, p.174.

[932] Ofcom, 2023. November 2023 Consultation, Volume 4, p.174.

[933] Section 21(4)(b) and Section 32(4)(b) of the Act.

content that may otherwise be missed. This can result in fewer users encountering illegal content, reducing the harm that may be caused when such content is allowed to circulate online.

6.39 Removal of illegal content can also reassure users that their complaints are taken seriously, leading to greater trust in (and increased use of) reporting and complaints procedures.

6.40 By recommending that service providers set up their complaints systems and processes to enable prospective complainants to submit all kinds of relevant complaints, this measure enables providers to process and take appropriate action against all kinds of complaints that the Act refers to.

## Costs and risks

6.41 Sections 21(3) and 32(3) of the Act require all service providers to handle relevant complaints. Given the wording of the measure closely follows these requirements – and allows discretion on how to achieve what is required – we consider its costs to be necessary to achieve what is required by the Act. We have therefore not sought to quantify them.

6.42 We recognise that additional costs may be incurred by operating more than one reporting and complaints process or system. However, we consider that this is required by the Act. Providers can decide the most appropriate approach for their services, giving them control and flexibility over the set-up and operating costs of their reporting and complaints processes. This flexibility will allow services to take an approach proportionate to their level of risk and their capacity while meeting the Act's requirements.

## Rights impact

6.43 As a result of a complaint, a service provider may take steps that (negatively or positively) affect the rights of users and others to freedom of expression, freedom of association, and privacy.[934]

### Freedom of expression and freedom of association

6.44 As explained in 'Introduction, our duties, and navigating the statement', Article 10 of the ECHR sets out the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 11 of the ECHR sets out the right to associate with others. We must exercise our duties under the Act in light of users', interested persons' and services' Article 10 and 11 rights and not interfere with these rights unless we are satisfied that to do so is prescribed by law, pursues a legitimate aim, is proportionate to the legitimate aim and corresponds to a pressing social need.

6.45 This measure may positively affect rights to freedom of expression and association. For example, a process for raising complaints with the service about illegal content could result in more effective moderation, creating safer spaces online. Users may feel more able to join online communities and share ideas and information with other users while being safeguarded from potential harm.

6.46 Potential interference with the rights to freedom of expression and association may arise where the service provider decides, as a result of a complaint, to restrict access to material it considers to be illegal content, or restricts users' ability to use the service (for example, by

[934] Articles 8, 9, 10 and 11 of the European Convention on Human Rights (ECHR).

banning or suspending them) on the basis of incorrect assessments of the nature of the content. We consider that the impact on users can be significantly mitigated by having a mechanism for appealing against incorrect decisions. We consider these matters in more detail in relation to our measures on content or search moderation (ICU C1, ICS C1).

6.47 We consider the impact of this measure to be relatively limited. It is likely to constitute the minimum degree of interference required to ensure that the service provider fulfils its duties under the Act.

#### Privacy

6.48 All complaints systems and processes will involve the processing of personal data of individuals, including children and those who are not users of the service, such as affected or interested persons. It will therefore affect users' rights to privacy and their rights under data protection law. The impact on users' or other individuals' rights would also be affected by the nature of the action taken as a result of the complaints process

6.49 As explained in our 'Introduction, our duties, and navigating the statement' Article 8 of the ECHR confers the right to respect for individuals' private and family life. An interference with the right to privacy must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest. In order to be 'necessary', the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.

6.50 The duty for a service provider to operate complaints systems and processes that enable relevant complaints is a requirement of the Act. This measure gives service providers flexibility as to precisely how they enact this requirement and what action they take. We recognise that the impact of the measure on users' privacy will depend on how a service provider decides to implement it. However, as noted in paragraph 6.29, it remains open to the service provider to decide how to operate their complaints procedure and to decide what forms of personal data they consider necessary to gather to process complaints, so long as they comply with the Act and the requirements of data protection legislation.

6.51 We consider that the privacy impact of this measure on prospective complainants will be relatively limited. We consider the impacts of the processing which follows receipt of a complaint in more detail in relation to our measures on content or search moderation (ICU C1, ICS C1) and appropriate action in response to complaints (ICU D7-ICU D12 and ICS D6-ICS D11). Assuming service providers comply with data protection legislation requirements, this measure is likely to constitute the minimum degree of interference required to secure that service providers fulfil their illegal content safety duties under the Act.

### Who this measure applies to

6.52 As set out in the Act, all providers of services – whether U2U or search services – must enable complaints systems and processes for prospective complainants to make all types of relevant complaints. This measure therefore applies to all U2U and search services.

## Conclusion

6.53 Enabling users to make relevant complaints is a requirement of the Act and will help service providers take appropriate action in relation to complaint, thereby reducing the risk of harm and the potential exposure of users to illegal content.

6.54 We are including this measure in our Illegal Content Codes of Practice for U2U and search services and is referred to within these Codes as ICU D1 for U2U services and ICS D1 for search services. It is part of our Codes on terrorism, CSEA and other duties.

## Components of the measure on easy to find, easy to access and easy to use complaints systems and processes

6.55 In our November 2023 Consultation, we proposed that all providers of U2U and search services should offer a complaints system for all types of relevant complaints that is easy to find, access, and use.[935] This directly addresses duties in the Act concerning how a service provider should operate its complaints procedures.

6.56 This is a measure with several components that address different duties and objectives. Because of this, we consider the measure across two separate sections in this chapter.

6.57 In this section, we consider the first four components of this measure. These are aimed at creating complaints systems that are easy to find and easy to use. In our November 2023 Consultation, we proposed that:

- reporting functions for relevant complaints regarding specific content should be clearly accessible in relation to that content for U2U services and search content for search services,
- the process for making other kinds of relevant complaints should be easy to find and accessible,
- the complaints process should only include as few steps as reasonably practicable,
- UK users and affected persons (and interested persons for search services) should have the ability to give the provider relevant information and supporting material when making a relevant complaint.

### Summary of stakeholder feedback[936]

6.58 Overall, stakeholders were positive about most, if not all, the components of the measure discussed in this section.[937] We identified several themes from stakeholder responses, which largely revolved around the implementation and feasibility of some of the components:

- The location of reporting tools in relation to content.

[935] Ofcom, 2023. November 2023 Consultation, Volume 4, p.177.

[936] Note: this list is not exhaustive and further responses can be found in Annex 1.

[937] 5Rights Foundation response to November 2023 Consultation, p.23; Betting and Gaming Council response to November 2023 Consultation, p.9; Global Partners Digital response to November 2023 Consultation, pp.17-18; National Trading Standards Scams Team response to November 2023 Illegal Harms Consultation, pp.1-2; New Zealand Classification Office response to November 2023 Consultation, p.8; NSPCC response to November 2023 Consultation, p.31; NWG Network response to November 2023 Consultation, p.9; South West Grid for Learning (SWGfL) response to November 2023 Illegal Harms Consultation, p.15; Refuge response to November 2023 Consultation, p.17; Scottish Government response to November 2023 Illegal Harms Consultation, p.8; Spotify response to November 2023 Illegal Harms Consultation, p.18.

- The number of steps needed to make a complaint.
- Allowing users to submit supporting material.

6.59 We detail comments on these themes in the following paragraphs.

### The location of reporting tools in relation to content

6.60 Skyscanner expressed concerns about the useability of a search service where a reporting function was located next to each piece of search content, arguing that it could be an "overly prescriptive" approach.[938] We discuss this feedback in paragraph 6.69 in the 'How these components work' section and paragraphs 6.78-6.80 in the 'Benefits and effectiveness' section.

### The number of steps needed to make a complaint

6.61 Meta argued that requiring services to set up complaints procedures that have as few steps as reasonably practicable risked limiting a provider's flexibility to include additional steps that may be logical to a complaints process.[939] Google argued that a more appropriate metric would be the extent to which a reporting process was intelligible to users, rather than the number of clicks needed to make a report.[940] We address these concerns in paragraph 6.71 in the 'How these components work' section and paragraphs 6.81-6.82 in the 'Benefits and effectiveness' section.

### Allowing users to submit supporting material

6.62 Several stakeholders agreed with our proposal that complainants should be able to provide contextual information when making a complaint.[941] The Institute for Strategic Dialogue cited examples of cases where there is repeated or persistent harm, arguing that the ability to provide contextual information would offer services greater understanding of the user's experiences.[942] The National Society for the Prevention of Cruelty to Children (NSPCC) highlighted that this component in the measure could help expedite the process of removing harmful content.[943] Similarly, 5Rights Foundation and Global Partners Digital highlighted that the ability to provide additional context could be important for complaints where, in isolation, a piece of content may not meet the threshold of illegality or its illegality may be difficult to determine.[944]

6.63 UK Finance noted that automated systems often fail to understand users' intentions and called for more human involvement in reporting systems, including the need to allow for "adequate context".[945] Some stakeholders suggested that automated systems sometimes

[938] Skyscanner response to November 2023 Consultation, p.21.
[939] Meta response to November 2023 Consultation, p.28.
[940] Google response to November 2023 Illegal Harms Consultation, p.49.
[941] 5Rights Foundation response to November 2023 Consultation, p.23; Big Brother Watch response to November 2023 Illegal Harms Consultation, pp.9-10; Global Partners Digital response to November 2023 Consultation, p.17; Institute for Strategic Dialogue response to November 2023 Consultation, p.10; Match Group response to November 2023 Consultation, p.13; NSPCC response to November 2023 Consultation, p.31; Philippine Survivor Network response to November 2023 Illegal Harms Consultation, p.11; UK Finance response to November 2023 Illegal Harms Consultation, p.10.
[942] Institute for Strategic Dialogue response to November 2023 Consultation, p.10.
[943] NSPCC response to November 2023 Consultation, p.31.
[944] 5Rights Foundation response to November 2023 Consultation, p.23; Global Partners Digital response to November 2023 Consultation, p.25.
[945] UK Finance response to November 2023 Consultation, p.8.

overlook or struggle to establish the context of content.[946] Similarly, with reference to the Troubles in Northern Ireland, South East Fermanagh Foundation (SEFF) highlighted that service providers need to understand the context behind instances of abuse that users may be reporting.[947]

6.64 LinkedIn expressed concerns about the feasibility and the benefits of this component of the measure.[948] Google stated it was not always "necessary or proportionate to provide users with the ability to submit supporting material in addition to text-based information", not least because of technical burdens in building such functionalities for all complaints systems. It suggested amending the Codes to clarify that users should be able to submit further information when making complaints, "but only to an extent reasonably appropriate for the circumstances."[949]

6.65 We discuss these stakeholder responses in paragraph 6.73-6.74 in the 'How these components work' section and in paragraphs 6.83-6.87 in the 'Benefits and effectiveness' section.

## Our decision

6.66 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation. We have made some amendments to some of the components of the measure discussed in this section

- We have changed the component of the measure regarding the number of steps in a complaints process. We now state that a complaints process should only include reasonably necessary steps.[950]
- We have amended the component of the measure related to allowing users to submit supporting information. Service providers have flexibility over what kinds of information they allow prospective complainants to submit, so long as it is clear that they can submit such information and where they can do so.[951]

6.67 The full text of these components of the measure can be found in our Illegal Content Codes of Practice U2U and search services and these components are referred to within the Codes as ICU D2.2a-d for U2U services and ICS D2.2a-d for search services. The measure is part of our Illegal Content Codes on terrorism, CSEA and other duties.

## Our reasoning

### How these components work

6.68 These components relate to how service providers can meet their duties under the Act to have a transparent and easy-to-use complaints procedure. The four components discussed in this section specifically address how complaints systems and processes should be set up so that users can easily find and use them.

[946] 5Rights Foundation response to November 2023 Consultation, p.24; UKSIC response to November 2023 Consultation, pp.38-39.
[947] South East Fermanagh Foundation (SEFF) response to November 2023 Illegal Harms Consultation, pp.8-9.
[948] LinkedIn response to November 2023 Consultation, pp.13-14.
[949] Google response to November 2023 Consultation, p.52.
[950] In our Codes, the relevant component is ICU D2.2.c for U2U services and ICS D2.2.c for search services.
[951] In our Codes, the relevant component is ICU D2.2.d for U2U services and ICS D2.2.d for search services.

6.69 The first component of the measure is designed to make reporting and complaints systems and processes easy to find for users as per Sections 20(2) and 31(2) of the Act.[952] Our measure says that, for relevant complaints regarding a specific piece of content, a reporting tool or function should be clearly accessible in relation to that content. We acknowledge stakeholder feedback about how this might affect the interface or useability of a U2U or search service.[953] Displaying a reporting tool alongside each search result is one clear way to ensure complaints processes are easily accessible. However, this is not essential or feasible in all cases and our measure does not mandate this. We recognise there are other ways to ensure complaints systems and processes are accessible and easy to use that may be more appropriate for different types of interfaces on both search and U2U services. Therefore, we are not making any amendments to the relevant component as we consider that our recommended measure already provides sufficient flexibility to service providers. Providers have flexibility over the design of reporting systems and processes, but should design them in a way that ensures they are easy to find in relation to a piece of content that a prospective complainant suspects is illegal content. This is the outcome this component seeks to achieve.

6.70 The second component of the measure relates to how easy a complaints system or process is for prospective complainants to find and navigate for other kinds of relevant complaints.[954] It outlines how services can comply with the duties in the Act to enable relevant complaints other than reports about content, such as complaints about the operation of a reporting tool or function.

6.71 The third component of the measure is about the number of steps in a complaints procedure.[955] In response to our 2022 Illegal Harms Call for Evidence, TrustElevate recommended that reporting functions or tools should have "the minimum number of clicks and steps for a user to quickly submit a report or complaint with ease while equipping the receiving party/platform with sufficient information to assess the report and determine the appropriate response".[956] We acknowledge that there is a balance required between making the complaints system or process quick and easy to use and ensuring that the service provider has the necessary information to handle complaints. We also acknowledge other stakeholder feedback about how the number of steps does not necessarily equate to ease of use.[957] We have changed this component of the measure to clarify we are not prescriptive over the number of steps a complaints system or process should entail: it should only include reasonably necessary steps. This will contribute to achieving the intended outcome of this measure that a complaints procedure is easy to use.

6.72 The fourth component of the measure allows prospective complainants to submit supporting information to supplement their complaint.[958] This component is designed to ensure service providers are able to better understand the context of a complaint to make an accurate assessment.

[952] In our Codes, the relevant component is ICU D2.2.a for U2U services and ICS D2.2.a for search services.

[953] Skyscanner response to November 2023 Consultation, p.21.

[954] In our Codes, the relevant component is ICU D2.2.b for U2U services and ICS D2.2.b for search services.

[955] In our Codes, the relevant component is ICU D2.2.c for U2U services and ICS D2.2.c for search services.

[956] TrustElevate response to Ofcom 2022 Call for Evidence: First phase of online safety regulation, p.8.

[957] Google response to November 2023 Consultation, p.49; Meta response to November 2023 Consultation, p.28.

[958] In our Codes, the relevant component is ICU D2.2.d for U2U services and ICS D2.2.d for search services.

6.73 Taking stakeholder feedback into consideration, we have changed the language in our Codes to say that services should allow prospective complainants to submit "supporting information" to give service providers flexibility over how they implement this component. Service providers should determine what format that "supporting information" might take. In some cases, it may suffice to include an option to include text-based information as supporting information. Other providers may wish to include a broader range of information formats. Ultimately, a service provider must allow users to submit supporting information of some form, and should ensure that the form(s) of supporting information they allow is appropriate to the service and the type of complaints the provider receives.

6.74 Some users may find that their report or complaint is not as credible without supplementary context. Without the option to include supporting information, a prospective complainant may find it difficult to justify reporting content they suspect to be illegal. For complainants who need to be able to report multiple items of content, allowing them to submit supporting information may also be an aspect of making complaints systems easy to use.

## Benefits and effectiveness

6.75 The primary benefit of these components of the measure is that they will make it easier for prospective complainants to report content they suspect to be illegal. In turn, this will enable service providers to take appropriate action.

6.76 We consider each component discussed in the following paragraphs and explain why we consider that they will deliver important benefits. We also set out where changes have been made to the components in response to stakeholder feedback.

### Measure on reporting tools being easily accessible (ICU D2.2.a and ICS D2.2.a) and easy to find and accessible processes for making other complaints (ICU D2.2.b and ICS D2.2.b)

6.77 Our research suggests that making reporting systems and processes more prominent or visible increases the likelihood of users reporting content.[959] There is also precedent for such a recommendation: the Australian Social Media Online Safety Code states that providers should ensure that reporting tools are "visible and accessible at the point the Australian end-user accesses materials".[960]

6.78 If prospective complainants struggle to understand a complaints procedure or locate a reporting system or process in order to submit a complaint about content they suspect to be illegal, service providers will be less likely to be able to take appropriate action against it, increasing the likelihood of harm to other users who may be exposed to it. Systems and processes that are easy to use will reduce risks in these areas. We expect these components to encourage prospective complainants to more readily report content they suspect to be illegal.

6.79 Providers should ensure prospective complainants are able to report content by other means even if they are not registered on the service. We are not prescriptive about how service providers enable this. As explained in paragraph 6.30 in 'Measure on enabling

[959] Ofcom's research into the impact of behaviourally informed designs for content-reporting mechanisms for VSPs found that raising the prominence of the reporting function increased the likelihood of reporting legal but potentially harmful content and categorising it accurately, while not appearing to increase over-reporting of neutral content. Ofcom, 2023. Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour, p.6 [accessed 10 November 2024].

[960] Australian E-Safety Commissioner, 2023. Australian Social Media Online Safety Code, p17.

complaints', the Act states that a 'user' does not have to be registered with a service. The Act and our measures require all users to be able to submit complaints, which means that the status of the user in relation to the service should not be a barrier to submitting complaints.

6.80 We conclude that accessible tools that are easy to find and use – including clear and prominent text and icons located close to the content being viewed and clearly signposting a 'reporting' or 'complaint' function – will encourage users to submit complaints and improve service providers' awareness of (and action against) illegal content. By increasing the proportion of illegal content that is reported, detected, and removed, this measure will reduce the number of users who are exposed to illegal content. This will contribute significantly towards overall safety.

### Measure on reporting processes only including reasonably necessary steps (ICU D2.2.c and ICS D2.2.c)

6.81 Our aim with this component is to reinforce the accessibility of complaints procedures and ensure that prospective complainants are not deterred from submitting complaints, whether that is because it takes too long or because the process is not easy to follow. In developing its EAST framework for designing behavioural interventions, the Behavioural Insights Team ('BIT') stated that its own research and behavioural literature indicated that small, seemingly irrelevant details that make a task more challenging could make a difference between an individual doing something and putting it off. An important principle to consider, therefore, is how to make it easier for someone to do something.[961] In our 2024 research into children's experiences of violent content online, we noted that children who had reported content had experienced issues with how long it took to submit the complaint and felt that the process was too complex and 'designed for adults'.[962]

6.82 Several stakeholders argued that the number of steps in a reporting and complaints procedure did not necessarily correlate with its accessibility. We recognise this and have made a small change to the wording of this component. It now recommends that complaints procedures should only include reasonably necessary steps. This will allow services to introduce additional steps where they make the complaints procedure simpler to navigate, whilst ensuring that steps are not overly complicated or excessive in length. It will also ensure providers have the flexibility to design their complaints systems and processes in a way that suits the needs of their service and users. We expect this to prove beneficial to all users – including children – by making complaints systems and processes easier to navigate and improving user experience.

### Measure on complainants having the ability to give supporting information (ICU D2.2.d and ICS D2.2.d)

6.83 The primary benefit of this component of the measure is that, in giving prospective complainants the option to submit supporting information when making a complaint, the likelihood of submitting successful complaints that result in harm-reducing decisions increases. We consider there to be several reasons for this.

[961] Service, O., Hallsworth, M., Halpern, D., Algate, F., Gallagher, R., Nguyen, S., Ruda, S., Sanders, M., 2015, EAST: Four simple ways to apply behavioural insights, pp.9-18 [accessed 10 November 2024].

[962] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children, p.37 [accessed 10 November 2024].

6.84 First, supporting information can provide important context to content a user suspects to be illegal. Context can help moderators to make an informed judgement about a complaint and correctly identify illegal content. In some instances, the lack of supporting information might result in valid complaints not being upheld due to missing context. For example, the Integrity Institute highlighted that complaints may be 'denied' without context.[963] Multiple respondents to our 2022 Illegal Harms Call for Evidence highlighted the same concern.[964] In response to our May 2024 Consultation, Epic Games described how it is currently implementing features that allow users to submit supporting information with a complaint.[965]

6.85 Second, prospective complainants may feel that their complaints require context to be considered appropriately (and subsequently deemed legitimate) and may consider not reporting if they are unable to provide additional context. As several stakeholders highlighted, for users submitting complaints which are appeals, being able to submit supporting information is an important way of establishing context and upholding rights.[966]

6.86 Third, the ability to provide additional context can make reporting and complaints systems and processes easier to use by reducing the burden of reporting. In response to our 2022 Illegal Harms Call for Evidence, Refuge highlighted that "survivors must usually report individual pieces of content in turn. Perpetrators will often send dozens or hundreds of messages, making reporting time-consuming and potentially [a] re-traumatising process for survivors".[967] The ability to provide supporting information (such as screenshots or descriptions showing how the user has been subjected to a pattern of behaviour, or the identities of the accounts engaging in the behaviour concerned) would reduce this burden on the user concerned. We consider this likely to be helpful for those who are at risk of harm from harassment and offline violence, many of whom are women and girls.

6.87 Supporting information is also important for users who are submitting complaints about a problem with a service provider's reporting tool or an instance of non-compliance with their safety duties. While our amendment to this component allows service providers to set the parameters regarding the type of supporting information they will accept, they should consider the needs of their users and should meet the outcome of creating an easy-to-use and accessible complaints procedure which enables them to make informed judgements.

## Costs and risks

### Costs

6.88 The components in this measure set out how we recommend services meet the specific requirements in the Act relating to complaints. We do not specify precisely how services

[963] Integrity Institute response to May 2024 Consultation on Protecting Children from Harms Online, pp.9-10.
[964] The Antisemitism Policy Trust cited an instance where it had reported a picture that, with additional context, allowed it to demonstrate an instance of far-right stalking of a high-profile Jewish individual. Refuge provided an example of survivors of domestic abuse having received images of their front doors and road signs after moving to a new location. Without context, an image of a front door is not harmful in itself and so is unlikely to be removed by content moderators; with added context, it may be reasonable to infer that the content amounts to harassment. Antisemitism Policy Trust response to Ofcom 2022 Call for Evidence: First phase of online safety regulation, p.10; Refuge response to Ofcom 2022 Call for Evidence: First phase of online safety regulation, pp.3-4.
[965] Epic Games response to May 2024 Consultation on Protecting Children from Harms Online, pp.15-16.
[966] Big Brother Watch response to November 2023 Consultation, pp.9-10; NSPCC response to November 2023 Consultation, p.31.
[967] Refuge response to Ofcom 2022 Call for Evidence: First phase of online safety regulation, pp.7-8.

should design their complaints systems and processes, and instead set out high-level recommendations leaving wide discretion to the service provider on how to achieve what is required. Most of the costs of the components therefore relate to the specific requirements in the Act, over which we have no discretion.

6.89 Possible costs related to the specific requirements of the Act include:

- one-off implementation costs for designing required changes,
- engineering costs of testing and implementing those changes, and
- costs of further refining the complaints system to ensure it continues to meet requirements over time.

6.90 While the Act does not specifically require that complainants should be able to provide supporting information when submitting a complaint, a complaints system must be "easy to access" and "easy to use". As explained in paragraph 6.72, we consider that allowing users to submit supporting information is a way of making complaints systems easy to use for some complainants. Other types of illegal harms (such as fraud, other harms committed using coded language and offences requiring a significant offline element) can be difficult to identify from a single item of content without extra information. It is therefore hard to see how a provider could keep its users safe from them if it did not consider support information. The setting up and operation of such a functionality may result in additional costs to providers, but we consider that these costs can be managed. As service providers have discretion over what types of information can be submitted, they can choose what is most appropriate for their service

### Risks

6.91 There may be a risk that the increased ease of use of complaints systems and processes could lead to an increase in false or malicious complaints. This could result in increased costs for the service, and could also lead to impacts on rights to freedom of expression or association. Overall, we consider both these risks to be manageable. Our measure is designed to give service providers flexibility and help them manage the costs of operating robust complaints systems and processes. This includes recommending that all relevant complaints are determined via a service provider's prioritisation and content or search moderation processes. Assuming these systems and processes are working effectively, we consider it feasible for a service provider to manage the costs of this risk, and unlikely that a disproportionate number of false or malicious complaints will be wrongly upheld.

6.92 Our measure recommends that users have the option to include supporting information when submitting a complaint, but we have not been prescriptive over what form 'supporting information' should take. We recognise that there is a risk that this flexibility might result in a service provider limiting the amount of information that users can submit, resulting in a user being unable to provide sufficient context for their complaint. We expect providers to find a way that allows complainants to give sufficient context for their complaint. We have not been more prescriptive about how this is done because the most appropriate approach will vary among services.

## Rights impact

### Freedom of expression and freedom of association

6.93 We consider that these components of the measure have the potential to affect prospective complainants and others' rights to freedom of expression and to freedom of association for similar reasons to those set out in paragraphs 6.44-6.47 under the 'Rights impact' section in 'Measure on enabling complaints'. We also consider the likely degree of interference with these rights to be limited for the reasons set out in paragraphs 6.46-6.47.

6.94 We consider that allowing prospective complainants to submit relevant information or supporting material could positively benefit their rights, particularly where they might have had their access to the service restricted or where access to content they have uploaded is restricted for other users on the basis that it is illegal content.

6.95 We consider that the impact of the components considered in this section on prospective complainants' or others' rights to freedom of expression and of association is limited and is likely to constitute the minimum degree of interference required to ensure that service providers fulfil their illegal content safety duties under the Act.

#### Privacy

6.96 The components considered in this section could affect prospective complainants' and others' right to privacy for the reasons set out in paragraphs 6.48-6.51 under the 'Rights impact' section in 'Measure on Enabling Complaints'.

6.97 Our recommendation that providers allow the submission of relevant information or supporting material when making a complaint may affect a user's right to privacy, whether they be a complainant or otherwise. Privacy impacts will depend somewhat on the extent to which the nature of any affected content is public or private (or, in other words, gives rise to a legitimate expectation of privacy). However, a prospective complainant can decide what relevant information or supporting material they share, and there should be no obligation on complainants to include personal information within these submissions. Where any additional personal data is provided, including data relating to individuals who are the subject of a complaint (such as profile information), it must be handled in accordance with data protection laws.

#### Data protection

6.98 As all complaints procedures will necessitate the collection, processing, and storing of personal data, the submission of a complaint may affect prospective complainants' rights to privacy and their rights under data protection law.

6.99 We consider that the impact of the components considered in this section on prospective complainants' and others' rights to privacy will be relatively limited and is likely to constitute the minimum degree of interference required to ensure that service providers fulfil their duties under the Act.

### Who these components apply to

6.100 Under sections 21 and 32 of the Act, all service providers are required to have complaints procedures that are easy to access, easy to use (including by children), and transparent. These components of the measures are therefore applicable to all providers of U2U and search services. The detail of these components gives service providers some discretion over how to apply it to suit their contexts.

## Conclusion

6.101 The Act requires that complaints procedures must be easy to access and easy to use. We consider the components discussed to be a proportionate way of achieving this. While some components, such as requiring complainants to be able to provide relevant information or supporting materials, are not explicitly required by the Act, our analysis suggests that they will all contribute in important ways to the goal of making complaints procedures easy to use. We therefore consider them a proportionate way of achieving the requirements in the Act, especially given we are allowing providers some flexibility over how they do this.

6.102 As explained above, we consider these components will deliver benefits to prospective complainants and service providers, and that the resulting costs to service providers and the impacts on prospective complainants' rights are proportionate. Complaints systems and processes that are easy to find, access, and use will enable prospective complainants to report content more easily.

6.103 These components are slightly amended from what we proposed in our November 2023 Consultation. Our changes to the measure are in response to stakeholder feedback. They allow providers more flexibility to design complaints procedures in a way that is appropriate for their services while still meeting the Act's requirements.

6.104 We are including the components of this measure in our Illegal Content Codes of Practice for U2U and search services, in which they are referred to as ICU D2.2a-d for U2U services and ICS D2.2a-d for search services. The components of this measure are part of our Illegal Content Codes on terrorism, CSEA and other duties.

## Components of the measure on accessibility of complaints systems

6.105 In our November 2023 Consultation, we proposed that all providers of U2U and search services should offer a complaints system for all types of relevant complaints that is easy to find, access, and use.

6.106 In this section, we consider the remaining three components of this measure that concern the accessibility needs of a UK userbase when using a complaints procedure or system. In our November 2023 Consultation, we proposed that:

- When designing their complaints processes, service providers should have regard to the particular needs of their UK userbase as identified in their risk assessment, and the needs of children and disabled users.
- Written information should be comprehensible based on the likely reading age of the youngest person permitted to use the service. In our May 2024 Consultation, we updated this to say that this should be based on the likely reading age of the youngest person permitted to use the service *without the consent of a parent or guardian*.
- The process should be designed to ensure usability for users with a disability or other accessibility needs (such as users of assistive technologies including keyboard navigation and screen reading tools).

6.107 These components of the measure aim to ensure that all reporting and complaints processes are accessible to the greatest possible number of users. This meets the duties set out in the Act requiring service providers to operate a complaints procedure that is 'easy to access, easy to use (including by children) and transparent' (Sections 21(2)(c) and 32(2)(c).

6.108 We have interpreted 'access' broadly to incorporate accessibility requirements and needs. While the Act sets out that we should consider accessibility from the perspective of children, we consider accessibility from the perspective of disabled people and vulnerable or at-risk groups to be equally important here.

6.109 This measure was mostly the same for both U2U and search services, with the only exception being that the measure also applies to UK 'interested persons' for search services, in addition to UK users and affected persons.

## Summary of stakeholder feedback[968]

6.110 Stakeholders welcomed our intention with these components of the measure and commented on our proposed recommendations about the accessibility of complaints procedures.[969] We received feedback from stakeholders about how service providers should make their complaints function accessible:

- User needs (including comprehension, disability and age).
- Choice of language in which to submit complaints.

6.111 We detail comments on these themes in the following paragraphs.

### User needs (including comprehension, disability and age)

6.112 One stakeholder called for more guidance on how we recommend providers assess needs and accessibility requirements, especially on ensuring comprehension based on the age of the youngest user.[970] Another stakeholder emphasised the need for easy to find complaints procedures that were easy to understand, but highlighted the need for service providers to invest in user interface design that would facilitate this.[971] Other stakeholders had similar concerns and suggestions regarding accessibility.[972] We address this feedback in paragraphs 6.117-120 in the 'How these components work' section.

6.113 Meta suggested that transparency may be reduced if information is designed to be comprehensible and accessible for children.[973] We consider this in paragraph 6.119 in the 'How these components work' section.

### Choice of language in which to submit complaints

6.114 Several stakeholders called for complaints processes to be offered in languages other than English, or for language requirements to be considered in the design of processes.[974] We

[968] Note: this list is not exhaustive and further responses can be found in Annex 1.

[969] 5Rights Foundation response to November 2023 Consultation, pp.23-24; ACNI response to November 2023 Consultation, p.9; Global Partners Digital response to November 2023 Consultation, pp.17-18; National Trading Standards Scams Team response to November 2023 Consultation, pp.1-2; New Zealand Classification Office response to November 2023 Consultation, pp.8-9; Refuge response to November 2023 Consultation, p.17; Scottish Government response to November 2023 Consultation, p.8; The Cyber Helpline response to November 2023 Consultation, p.15; Yoti response to November 2023 Illegal Harms Consultation, p.14.

[970] Yoti response to November 2023 Consultation, p.14.

[971] INVIVIA response to November 2023 Consultation, p.18.

[972] Global Partners Digital response to November 2023 Consultation, pp.17-18; Internet Matters response to November 2023 Illegal Harms Consultation, p.17; Scottish Government response to November 2023 Consultation, p.8; The Cyber Helpline response to November 2023 Consultation, p.15.

[973] Meta response to May 2024 Consultation on Protecting Children from Harms Online, p.25.

[974] Centre for Competition Policy response to November 2023 Consultation, p.17; Global Partners Digital response to November 2023 Consultation, p.17; Open Rights Group response to November 2023 Consultation, p.2; Philippine Survivor Network response to November 2023 Consultation, p.11; Refuge response to November 2023 Consultation, p.17.

interpret this concern as relating to accessibility and address it in paragraph 6.122 in the 'How these components works' section.

# Our decision

6.115 We have decided to broadly confirm the measure we proposed in the May 2024 Consultation. We have made some changes to the components discussed in this section.

- We now recommend service providers consider the likely accessibility needs of their UK userbase, having regard to: (1) relevant information it holds on its UK userbase (including from its risk assessment and also from its children's risk assessment, if it was required to undertake one); (2) industry standards and good practice on accessibility for disabled people; and (3) comprehensibility based on the likely reading age of the youngest individual permitted to use the service without the consent of a parent or guardian.[975]

6.116 The full text of these components of the measure can be found in our Illegal Content Codes of Practice for U2U and search services on terrorism, CSEA and other duties. These components of the measure are referred to within these Codes as ICU D2.3-2.4 for U2U services and ICS D2.3-2.4 for search services.

# Our reasoning

## How these components work

6.117 Each service will have a different userbase with different needs and requirements that the service provider should make efforts to understand. Based on stakeholder feedback and our independent analysis, we have made some amendments to these components to establish a baseline from which service providers should understand 'accessibility', while allowing them the flexibility to make choices that suit the requirements of their services and userbase.

6.118 Service providers should consider the needs of their userbase when designing complaints systems and processes. We recommend a service provider does this by considering information it holds as a result of its risk assessment and children's risk assessment (where a service is required to undertake one). These assessments will help inform a provider's understanding of the demographics of its userbase, such as the age groups of users (including children) on its service, so that it can design its complaints procedures based on who is most likely to be using them.

6.119 Written information for users should be "comprehensible for the youngest individual permitted to use the service without the consent of a parent or guardian".[976] This is aligned with our recommendations for terms of service and publicly available statements and works to ensure that the greatest number of users possible – including adults with learning difficulties or disabilities – are able to understand complaints systems and processes. Most providers will present instructions, guidance and steps for complaints systems and processes in written format; if this written information is incomprehensible for the vast majority of users, the complaints procedure may be under-used or used incorrectly. With regard to stakeholder feedback about a reduction in transparency if information is being

[975] In our Codes, the relevant component is ICU D2.3 for U2U services and ICS D2.3 for search services.

[976] In our Codes, this measure is ICU G3/ICS G3 for Terms of Service and Publicly Available Statements.

designed with comprehensibility in mind, we consider that this expectation does not require a provider to reduce the amount of information it is sharing.[977] Instead, it ensures the users of the service are able to comprehend the information that is being communicated to them.

6.120 We also recommend that service providers consider industry standards and good practice when designing their complaint procedures so that they are appropriate to the access needs of disabled people. Industry standards and 'good practice' will look different for different services, but there are a range of techniques which can be effective. Some users with visual or motor impairment may depend on a keyboard to navigate webpages and functions or tools on a service. Others may require screen readers to make content on a screen accessible for those who are unable to see it or rely on the use of assistive technologies. Complaints systems and processes should be designed in a way that does not inhibit users with such requirements from navigating them. The World Wide Web Consortium's (W3C) Web Content Accessibility Guidelines are widely used throughout the industry and provide guidance and information on how services can be made more accessible.[978]

6.121 We consider that, in meeting accessibility needs, a service provider is better placed to facilitate accurate and appropriate reporting and complaints. Whilst these amendments represent a slight expansion of the original measure, they are still flexible enough for providers to make their own determinations on how best to design their complaints processes for their users.

6.122 Some stakeholders called for complaints systems and processes to be offered in languages other than English.[979] We have not been prescriptive about the language in which service providers should allow prospective complainants to make complaints. In requiring service providers to consider accessibility needs having regard to their risk assessments, we are expecting services to be able to determine which language(s) would best suit their userbase. For example, if a service's UK userbase consists of a significant proportion of non-English speakers, we would expect the service to design its complaints procedure in a language other than English to suit the needs of those users. Service providers have discretion over offering users an option of an alternative language in which to submit complaints but should ensure that the languages available are offered with the intention to suit the needs of their userbase.

### Benefits and effectiveness

6.123 Making all complaints systems and processes accessible helps to encourage prospective complainants to submit reports.

6.124 The accessibility of complaints systems and processes is particularly important to vulnerable users. This might include children, disabled people, or at-risk groups, all of whom often face additional barriers to reporting or submitting complaints. Making complaints systems and

[977] Meta response to May 2024 Consultation on Protection Children from Harms Online, p.25.

[978] World Wide Web Consortium (W3C), 2024. Web Content Accessibility Guidelines 2 Overview. [Accessed 24 October 2024].

[979] Centre for Competition Policy response to November 2023 Consultation, p.17; Global Partners Digital response to November 2023 Consultation, p.17; Open Rights Group response to November 2023 Consultation, p.2; Philippine Survivor Network response to November 2023 Consultation, p.11; Refuge response to November 2023 Consultation, p.17.

processes more accessible and easier to use will increase the likelihood that users, especially those with a greater chance of being exposed to particular illegal content, will submit complaints. This is the outcome we are seeking to achieve with these components in this measure.

6.125 We acknowledge that service providers – and the systems they operate – cannot guarantee that every user will find it easy to report or submit a complaint about content. However, consideration of users' needs (including those of children and disabled people) should help service providers create an inclusive system that will be easy to use for the largest number of users.

### Costs and risks

6.126 The accessibility components of these measures set out how we recommend service providers meet the specific access requirements in the Act relating to complaints. As with the components discussed in the preceding section, we do not specify precisely how services should design their complaints systems. Instead, we have left wide discretion to service providers on how to achieve the outcome required by the Act.

6.127 Possible costs related to the specific requirement of the Act include:

- one-off implementation costs for designing required changes;
- engineering costs of testing and implementing those changes; and
- costs of further refining the complaints system to ensure it continues to meet requirements over time.

6.128 We acknowledge that these components may increase ongoing costs for some service providers, though we expect any increases to be small. Most of these costs relate to the specific requirements in the Act to ensure that complaints procedures are accessible and easy to use, over which we have no discretion.

### Rights impact

6.129 We consider the rights impact of these component – including in reference to freedom of expression and association, privacy, and data protection – in the previous 'Rights impact' section under the previous measure (see paragraphs 6.93-6.99).

### Who these components apply to

6.130 Under sections 21 and 32 of the Act, all service providers are required to have complaints systems and processes that are easy to access, easy to use (including by children), and transparent. The measure containing these components are therefore applicable to all providers of U2U and search services.

## Conclusion

6.131 We consider that these components will deliver be beneficial to users. Considering accessibility in complaints systems and processes makes them easier to use, reducing the burden of reporting on users and ensuring service providers receive more accurate complaints. Given the importance of ensuring complaints procedures are accessible and the fact that most of the costs the measure imposes relate to specific requirements of the Act, we consider that this is a proportionate measure to include in Codes.

6.132 These components are slightly amended from what we proposed in our November 2023 Consultation and subsequently in our May 2024 Consultation. Our changes to the components allow providers more flexibility to design complaints procedures in a way that is appropriate for their services, while still meeting the Act's requirements.

6.133 We are including the components of this measure in our Illegal Content Codes of Practice for U2U and search services, and they are referred to as ICU D2.3-2.4 for U2U services and ICS D2.3-2.5 for search services. The components of this measure are part of our Codes on terrorism, CSEA and other duties.

## Measure on providing information prior to the submission of a complaint

6.134 In our May 2024 Consultation, we consulted on including a new measure in the Codes for U2U services to meet the needs of children.[980]

6.135 The measure recommended that a service provider ensures that complainants can easily access information on whether the provider will share information about a complaint with another user and what information will be shared. This also includes information about the original complaint and complainant if the other user subsequently appeals.

6.136 We proposed adding this measure to both the Illegal Content Codes and the Children's Safety Codes, and that it should apply to all U2U services likely to be accessed by children. The evidence we have suggests child users would benefit most from understanding whether their identities are disclosed or not after the submission of a complaint. We do not have such evidence for adult users. Our duties also require us to consider a higher level of protection for children than for adults.[981]

6.137 This measure is not a direct requirement of the Act but contributes to making complaints procedures transparent (Section 21(2)(c)), in particular for children, and to keeping children safe.

### Summary of stakeholder feedback[982]

6.138 Stakeholders were largely supportive of the objective of this measure.[983] We identified the following themes from stakeholder responses to our May 2024 Consultation:

- Anonymity.
- Application of this measure and the potential burden on services.

[980] In our May 2024 Consultation, this measure was listed as 'Measure PCU C3'.

[981] Section 3(4A)(b) of the Communications Act, 2003 sets out the 'need for a higher level of protection for children than adults'. We must also ensure that measures in our Codes of Practice are compatible with pursuit of the objective that the service is designed and operated in such a way that it 'provides a higher standard of protection for children than for adults', as per Schedule 4, para 4(a)(vi) of the Act.

[982] Note: this list is not exhaustive and further responses can be found in Annex 1.

[983] Commissioner Designate for Victims of Crime Northern Ireland response to May 2024 Consultation on Protecting Children from Harms Online p.6; Global Partners Digital response to November 2023 Consultation, p.16; NSPCC response to May 2024 Consultation on Protecting Children from Harms Online p.58; The Centre for Excellence for Children's Care and Protection (CELCIS) response to May 2024 Consultation on Protecting Children from Harms Online, p.15.

6.139 We detail comments on these themes in the following paragraphs.

### Anonymity

6.140 The NSPCC and the Centre for Excellence for Children's Care and Protection (CELCIS) welcomed the measure but highlighted that children may not feel comfortable submitting complaints without a guarantee of anonymity.[984] An individual respondent supported this, calling for anonymity to also be guaranteed at appeal stage.[985] Another stakeholder suggested that children may be reluctant to report for a myriad of reasons relating to their social lives and to concerns about how the complaints might impact them (for example, they may be concerned that they will be considered complicit in illegal activity).[986] We consider this feedback in paragraphs 6.147-6.149 in the 'How this measure works' section.

### Application of this measure and the potential burden on service

6.141 Snap suggested that this measure should be extended to all U2U and search services in scope of the Act. It highlighted that a variety of groups – beyond children – under-report due to concerns about anonymity in the reporting process.[987] Mid Size Platform Group said this measure could impose disproportionately high resource demands on service providers.[988] We respond to this feedback in paragraphs 6.156-6.159 in the 'Who this measure applies to' section.

## Our decision

6.142 We have decided to broadly confirm the measure we proposed in our May 2024 Consultation. We have made some clarificatory amendments and have changed who this measure applies to. The measure will apply only to providers of U2U services likely to be accessed by children that is either:

- a large service, or
- at medium or high risk for any kind of illegal harm.

6.143 The full text of the measure can be found in our Illegal Content Codes of Practice for U2U services and is referred to within these as ICU D3. It is part of our Codes on terrorism, CSEA and other duties.

## Our reasoning

### How this measure works

6.144 This measure increases transparency between users and services likely to be accessed by children by clarifying how information regarding a complaint and the complainant will be used or disclosed, if at all. The measure sets out that service providers should make sure this information is accessible *prior* to a complainant submitting a complaint. Depending on the way their complaints procedure is set up, providers may wish to display this information as part of the complaints process itself (for example, behind a question mark or help

[984] CELCIS response to May 2024 Consultation, p.15; NSPCC response to May 2024 Consultation, p.58.
[985] Dean, J. response to May 2024 Consultation on Protecting Children from Harms Online, p.16.
[986] Fuller, A. response to November 2023 Consultation, p.19.
[987] Snap response to May 2024 Consultation on Protecting Children from Harms Online, p.29.
[988] Mid Size Platform Group response to May 2024 Consultation on Protecting Children from Harms Online, p.11.

button) or a hyperlink to it. Given the wide range of service providers to which this measure applies, we do not consider it appropriate to be prescriptive about where exactly this information is located. However, we recommend that prospective complainants are able to access this information easily before submitting a complaint.

6.145 This measure does not make recommendations about exactly what information service providers should share with users other than the complainant. Instead, it recommends providers make it clear to complainants what information (if any) will be shared with any other users (including the user whose content is the subject of a complaint).

6.146 All complaints procedures will involve the processing of personal data. As such, they are subject to the requirements of the UK's data protection regime. This includes a requirement for services to put in place appropriate technical and organisational measures to implement data protection principles effectively and safeguard individual rights. Providers should consult ICO guidance (including the ICO's Children's Codes) to ensure that their complaints procedures protect user privacy in line with the data protection regime.[989]

6.147 We recognise that some children may not understand how their personal information is used, which may result in them under-using complaints procedures. However, given the protections and safeguards which are already in place under data protection laws, we do not consider it appropriate or proportionate to recommend in our Codes that service providers guarantee children's anonymity in complaints. Under data protection law, service providers are required to ensure that their processing of personal data is limited to what is necessary to achieve their objective, and we understand that it is not industry practice to share such information with other users.

6.148 There may sometimes be legitimate reasons why providers may need to know the identity of the complainant, for example to make a safeguarding or welfare referral. It may also sometimes be impossible for providers to prevent other users from working out that they were complained about and by whom through, for instance, a process of elimination or where the content was shared only with one other user.

6.149 Services should take the necessary precautionary steps to make sure that complainants are not identified inadvertently by, for example, sharing unique details about a complaint that enables the subject of that complaint to identify who the complainant is. Providers will need to make their own assessment of what personal information is necessary to process complaints.

## Benefits and effectiveness

6.150 This measure is designed to aid transparency in reporting and complaints, (which is also a duty outlined in the Act) by making it clear to users of a service what information is shared with whom before the submission of a complaint, and lift barriers to submitting a complaint. Our 2024 research into children's attitudes to reporting found that concerns about the disclosure of identity often deterred children from submitting complaints. Participants said they did not believe the reporting process to be anonymous and expressed

[989] Information Commissioner's Office, 2023. Data protection by design and default [accessed 10 November 2024]; Information Commissioner's Office, 2022. Age appropriate design: a code of practice for online services [accessed 10 November 2024].

concerns that their details would be included in a notification sent to the user they reported.[990]

6.151 Children participating in our research into experiences of cyberbullying said that anonymity was important to them when reporting to ensure that a report or outcome of a report could not be traced back to them, due to concerns that acting against a bully might exacerbate the situation.[991] Stakeholders supported this, explaining that being more transparent with children about what will happen when they submit a complaint could help increase their trust in complaints processes and make them more likely to complain about illegal and harmful content.[992]

6.152 Making this information easily accessible could also improve transparency for other prospective complainants by informing them of whether they should expect to be notified if their content or account is complained about. This explanation should be easily accessible.

## Costs and risks

6.153 The costs of this measure will vary depending on how providers choose to implement it. For example.

- some providers may choose to develop an interstitial or banner with this information accessible from the reporting tool, before the complaint is submitted,
- some providers may prefer to display or link to the information from the screen where a prospective complainant can submit a complaint, before the complaint is submitted.

6.154 We expect associated costs to largely be incurred in design, quality assurance (QA), and testing. As described in the May 2024 Consultation, we estimate the direct cost of implementing this measure to be approximately one day to two weeks of software engineering time (along with an equivalent amount of time input from professional occupational staff). Using our assumptions on labour costs required for this type of work set out in Annex 5, we expect the one-off direct costs to be approximately £400 to £9,000. We assume annual maintenance costs to be 25% of initial set-up costs and estimate these to be approximately £100 to £2,250 per annum. We expect providers of smaller services with simpler complaints procedures to incur costs towards the lower end of this range as they can deploy a simpler approach in making this information available as described in the previous paragraph.

## Rights impact

### Freedom of expression, freedom of association and privacy

6.155 We consider that this measure may have positive benefits on prospective complainants' rights to freedom of expression, freedom of association and privacy, and help to safeguard them. We do not expect this measure to give rise to any additional impacts on users' and others' rights.

[990] Ofcom, 2024. Understanding Pathways to Online Violent Content Among Children, pp.7,37 [accessed 10 November 2024].

[991] Ofcom, 2024. Key attributes and experiences of cyberbullying among children in the UK, p.45.

[992] Dean, J. response to May 2024 Consultation, p.16; NSPCC response to May 2024 Consultation, p.56.

### Who this measure applies to

6.156 In our May 2024 Consultation, we proposed to recommend this measure to providers of all U2U services that are accessed by children.

6.157 However, we have decided not to apply this measure to providers of small, low-risk services. As described in chapter 13 of this Volume: 'Combined impact assessment', several stakeholders expressed concerns about the potential impact of our measures as a whole on providers of smaller services that assess as low-risk for all kinds of illegal harms.

6.158 Most measures we recommend for providers of such services emerge from specific requirements in the Act over which we have no discretion. This is one of only a few measures we are recommending over which we have discretion and, after considering responses, we conclude that the benefits of this measure would be low for small, low-risk services.

6.159 Many providers of small, low-risk services are likely to receive few, if any, relevant complaints as there will not be a large amount of illegal content on their services. Therefore, we are not satisfied that it would be proportionate to apply this measure to these providers. There is also a risk that recommending this measure to providers of such services may harm users' interests because they may lose access to these services if providers withdraw them due to increased costs.[993]

6.160 This measure therefore applies to providers of all U2U services that are likely to be accessed by children, that are large or have a medium or high risk for any kind of illegal harm. Even if a provider only assesses its service as medium or high risk for a single harm, we consider this measure proportionate given its relatively low cost and the important benefits it will deliver. We maintain that it is beneficial to apply these measures to providers of large, low-risk services because they have the reach and potential to affect many users, as explained more fully in paragraph 1.156 of chapter 'Our approach to developing Codes measures'.

6.161 We are recommending this measure only for providers of U2U services that are likely to be accessed by children because the evidence we have on the benefits of this measure relates to children in particular. Our duties also require us to have regard to the need for a higher level of protection for children than for adults.[994]

6.162 We are not recommending this measure for search services because we do not have evidence that children had similar concerns about reporting content on search services.

## Conclusion

6.163 Our analysis shows that this measure can be beneficial to both users who are children and service providers. It can increase prospective complainants' trust in a complaints procedure, thereby increasing its use. It can aid service providers in ensuring their complaints procedures are transparent. Given that the measure will deliver important benefits and that

[993] This is consistent with our concerns on the overall burden on such services as discussed in Chapter 13 of this Volume: 'Combined impact assessment'.

[994] Section 3(4A)(b) of the Communications Act, 2003 sets out the 'need for a higher level of protection for children than adults'. We must also ensure that measures in our Codes of Practice are compatible with pursuit of the objective that the service is designed and operated in such a way that it 'provides a higher standard of protection for children than for adults', as per Schedule 4, para 4(a)(vi) of the Act.

the costs of the measure are relatively low, we consider it proportionate to include it in our Codes.

6.164 However, consistent with our risk-based approach, we have decided not to apply this measure to providers of services which are small and low risk. Given the low risk of harm occurring on these services, we consider it appropriate to minimise the regulatory burden on such services where possible. The measure will therefore only apply to U2U services that are likely to be accessed by children and are large or have a medium or high risk for any kind of illegal harm.

6.165 We are including this measure in our Illegal Content Codes of Practice for U2U services and is referred to as ICU D3. It is part of our Codes on terrorism, CSEA and other duties.

## Measure on acknowledging complaints and sending indicative timeframes

6.166 In our May 2024 Consultation, we proposed that all providers of U2U and search services should provide the complainant with an acknowledgement and an indicative timeframe for when their complaint might be decided. We had posed two options in our November 2023 Consultation:

- Option One: Complainants receive an acknowledgment of their complaint containing an indicative timeline for handling it.
- Option Two: The service provider takes a more detailed approach to enable complainants to check the status of their complaints or for updates to be proactively sent to users.

6.167 Our preferred option was to recommend that the service provider should acknowledge receipt of complaints with an indicative timeframe for deciding the complaint. We did not propose to recommend that the service provider necessarily needs to provide a contact person, progress updates, or information on the outcome of the complaint.

6.168 The provision of acknowledgements or an indicative timeframe is not a requirement set out by the Act, but it can incentivise users to submit complaints and service providers to deal with complaints appropriately and swiftly. It can also aid transparency, which is a requirement under the Act.

### Summary of stakeholder feedback[995]

6.169 Several stakeholders expressed their support for all or parts of this measure.[996]

6.170 Some stakeholders raised concerns about the measure, including the risk of unintended consequences. We have grouped these responses into themes:

[995] Note: this list is not exhaustive and further responses can be found in Annex 1.

[996] 5Rights Foundation response to November 2023 Consultation, pp.23-24; British and Irish Law, Education and Technology Association (BILETA) response to November 2023 Illegal Harms Consultation, pp.23-24; Centre for Competition Policy response to November 2023 Consultation, p.17; [✂]; Global Partners Digital response to November 2023 Consultation, p.16; ICO response to November 2023 Consultation, pp.18-19; INVIVIA response to November 2023 Consultation, p.18; Open Rights Group response to November 2023 Consultation, p.4; Snap response to November 2023 Consultation, p.14.

- Enabling users to track complaints.
- Safeguarding concerns related to services sending acknowledgements.
- Setting and meeting reasonable timeframes.
- Burden on services.

6.171 We detail comments on these themes in the following paragraphs.

## Enabling users to track complaints

6.172 Some stakeholders expressed a preference for the second option we considered in November 2023: that users should be able to track the progress of their complaints or kept informed of the status of their complaints. Pinterest said that transparency would be better served if users were able to track the status of their reports.[997] UK Interactive Entertainment ('Ukie') expressed that allowing users to check the status of their complaints would lead to a better user experience.[998] 5Rights Foundation, Global Partners Digital and Refuge suggested that users should be able to follow up or receive an update on their complaints, with Refuge and 5Rights Foundation considering this a way of considering a complainant's wellbeing.[999] We address this feedback in paragraph 6.183 in the 'How this measure works' section.

## Safeguarding concerns related to services sending acknowledgements

6.173 Glitch highlighted that women and girls may have specific preferences for how they receive information in response to their complaints due to safeguarding concerns.[1000] In response to our May 2024 Consultation, the NSPCC highlighted that an opt-out to acknowledgements could help prevent further distress.[1001] We consider this feedback in paragraph 6.187-6.190 in the 'How this measure works' section and paragraph 6.195 in the 'Benefits and effectiveness' section.

## Setting and meeting reasonable timeframes

6.174 Federation of Small Businesses and Refuge called for further guidance on what would be considered a reasonable timeframe, with Refuge suggesting Ofcom set minimum standards for timeframes to ensure clarity over the purpose of the measure.[1002] The ICO highlighted that whatever timeframe a service provider sets would need to consider time limits under data protection laws and other areas of law.[1003] We consider this feedback in the 'How this measure works' and 'Rights impact' sections.

6.175 Several stakeholders raised concerns about the feasibility of this measure given that response times often vary depending on the type of complaint and service.[1004] Vinted

[997] Pinterest response to November 2023 Illegal Harms Consultation, p.8.

[998] Ukie response to November 2023 Consultation, p.22.

[999] 5Rights Foundation response to November 2023 Consultation, p.24; Global Partners Digital response to November 2023 Consultation, pp.16-17; Refuge response to November 2023 Consultation, p.12.

[1000] Glitch response to November 2023 Illegal Harms Consultation, pp.9-10.

[1001] NSPCC response to May 2024 Consultation, p.57.

[1002] Federation of Small Businesses response to November 2023 Illegal Harms Consultation, p.3; Refuge response to November 2023 Consultation, pp.11-12.

[1003] ICO response to November 2023 Consultation, pp.18-19.

[1004] BILETA response to November 2023 Consultation, pp.23-24; Meta response to November 2023 Consultation, pp.28-29; Online Dating and Discovery Association response to November 2023 Illegal Harms highlighted that providing timelines could be administratively burdensome and difficult to assess at the point of acknowledging receipt of complaints.[1005] Meta further highlighted that this timeframe would also vary depending on the prioritisation the complaint is given.[1006] We address this feedback in the 'Benefits and effectiveness' section and in the 'Costs and risks' section.

6.176 Meta, Mega, and Snap also expressed concerns that this measure might undermine complaints procedures. Mega and Snap said that the measure might incentivise services to set longer timeframes to ensure that they could resolve complaints within that period.[1007] Meta said that users might get frustrated if complaints are not resolved within a given timeframe and that services might be incentivised to resolve complaints faster at the expense of accuracy.[1008] We address this feedback in paragraphs 6.201-6.203 in the 'Costs and risks' section.

### Burden on services

6.177 Two stakeholders expressed concerns that the measure would lead to a disproportionate burden on some services.[1009] Vinted considered the provision of timeframes 'excessive' relative to other measures and requirements.[1010] We address this feedback in paragraphs 6.182-6.185 in the 'How this measure works' section and paragraphs 6.198-6.200 in the 'Costs and risks' section.

## Our decision

6.178 We have decided to broadly confirm the measure we proposed in May 2024 Consultation, (based on our preferred option from the November 2023 Consultation), with some clarificatory amendments and two changes:

- we have amended the phrasing of the measure to refer to 'indicative timeframes' rather than 'indicative timelines'; and
- we have decided not to apply this measure to small services which are low risk for any kind of illegal harm. For U2U services, the measure will apply to a provider of a service that is either a large service, or at medium or high risk of any kind of illegal harm. For search services, the measure will apply to a provider of a service that is either a large general search service or at medium or high risk of any kind of illegal harm.

6.179 Reflecting on stakeholder feedback on this measure, we have decided to recommend service providers allow complainants to choose whether they receive communication from the provider after they have submitted a complaint.

6.180 The full text of the measures can be found in our Illegal Content Codes of Practice for U2U and search services on terrorism, CSEA and other duties. Within these, the measure is

Consultation, p.2; Pinterest response to November 2023 Consultation, p.8; Snap response to November 2023 Consultation, p.14; Vinted response to November 2023 Illegal Harms Consultation, p.12.

[1005] Vinted response to November 2023 Illegal Harms Consultation, p.12.

[1006] Meta response to November 2023 Consultation, pp.28-29.

[1007] Mega response to November 2023 Illegal Harms Consultation, p.5; Snap response to November 2023 Consultation, p.14.

[1008] Meta response to November 2023 Consultation, pp.28-29.

[1009] Mid Size Platform Group response to November 2023 Consultation, pp.9-10; Skyscanner response to November 2023 Consultation, p.21.

[1010] Vinted response to November 2023 Consultation, p.12.

referred to as ICU D4 (and see also ICU D6) for U2U services and ICS D3 (and see also ICU D5) for search services.

## Our reasoning

### How this measure works

6.181 The Act requires all service providers to operate complaints procedures that are transparent.[1011] We interpret this to mean transparency over how a service provider will handle complaints, including any immediate communication the service might have with complainants regarding how decisions are made and within what timeframes.

6.182 Our measure recommends service providers acknowledge complaints from complainants and provide indicative timeframes for deciding complaints. We do not stipulate a need for providers to direct prospective complainants to a contact person, offer progress updates, or provide information on the outcome of the complaint. As explained in our November 2023 Consultation, we do not have sufficient evidence of the practicalities and costs of implementing such a recommendation at scale for each of the types of complaints that providers are required to consider.[1012] Our measure gives the service provider flexibility over how they acknowledge a complaint. It can be automated as a 'pop-up', sent as an email, or take any other form a service provider feels appropriate.

6.183 We consider that, if a service provider has provided a complainant with an indicative timeframe within which their complaint will be considered, it has acknowledged the complaint. As such, an indicative timeframe can form part of a service provider's acknowledgement.

6.184 We recognise that timeframes for handling and resolving complaints will differ across service providers. As set out in paragraphs 6.174-6.176, several stakeholders highlighted that timeframes would vary considerably depending on the nature of the complaint, and could be administratively burdensome and difficult to assess at the point of acknowledging receipt of complaints.

6.185 We have amended this measure so that it recommends providers provide complaints with an indicative 'timeframe', which we consider broader and more flexible than an indicative 'timeline'. Our intent with this measure remains that these timeframes do not need to be bespoke to the specific complaint. This gives service providers flexibility over how they determine what is appropriate and realistic for them. Service providers can provide timeframes that are bespoke to the type of complaint, should they wish to.

#### Opting out of communications

6.186 We recognise stakeholder concerns about the safeguarding risks that this measure could give rise to. We are therefore recommending that service providers give complainants the option to opt out of receiving non-ephemeral communications in relation to a complaint.[1013]

6.187 While acknowledging complaints can lead to better reporting and complaints experiences, it can also have unintended consequences, especially for vulnerable users who may be in

[1011] Sections 21(2)(c) and 32(2)(c) of the Act.

[1012] Ofcom, 2023. November 2023 Consultation, Volume 4, p.189.

[1013] In our Codes, this measure is referred to as ICU D6 for U2U service and ICS D5 for search services.

abusive or coercive situations. Detailed analysis on the causes and experiences of coercive and controlling behaviour online can be found in the Register of Risks ('Register').[1014]

6.188 Given these concerns, we have to add a provision enabling complainants to opt out of receiving non-ephemeral communications. This means that complainants should be able to choose whether they receive an acknowledgement and indicative timeframe. Providers have flexibility over how to implement this opt-out. It could be done by incorporating a tick-box feature into existing reporting functions, where a complainant opts out of receiving an acknowledgement and the service does not send any further automated follow-ups.

6.189 The complainant would still be at risk, however, if a provider were allowing them to opt out from acknowledgements, but then sent other non-ephemeral communications about the complaint to them. We are therefore recommending that the opt-out apply to all non-ephemeral communications about the complaint. This means that service providers can offer complainants greater agency in tailoring their reporting and complaints experience in a way that is appropriate and safe for them.

6.190 Providers do not need to provide an opt-out option for ephemeral acknowledgements, i.e. acknowledgements that appear and disappear at the time the complaint is submitted and cannot be restored.

## Benefits and effectiveness

6.191 We envision three main benefits of this measure.

6.192 First, there is evidence that transparency over complaints procedures can encourage users to submit complaints about suspected illegal content. For example, in its 'Unsocial Spaces Report', Refuge highlighted that complainants often wait a long time to receive any information about their complaint and, in some cases, receive no response at all, which can compound their stress and trauma.[1015] Experiences in other sectors show that a response within two working days increased confidence in complaints handling.[1016] Lack of communication between a service provider and a complainant can also impact on reporting and complaining behaviours, leaving complainants to feel that their concerns are not being dealt with and reducing confidence in complaints procedures.

6.193 As outlined in our November 2023 Consultation, we consider that sending an acknowledgement and providing an indicative timeframe will signal to complainants that their complaints have been received and that appropriate action will be taken in response to them. We consider that if a service provider demonstrates transparency and communication to their userbase in this way, users will be more empowered to submit complaints about suspected illegal content. This is especially valuable for complaints about illegal content where a service provider may have to flag the content to law enforcement. For example, providers have a duty to flag CSEA to the National Crime Agency. We expect this to foster trust in the complaints process and increase the likelihood of it being used.

6.194 Second, this measure can incentivise service providers to deal with complaints appropriately and swiftly. Indicative timeframes - that the complainant is aware of - can

[1014] Register chapter titled 'Controlling or coercive behaviour'.

[1015] Refuge, 2021. Unsocial Spaces, p.25. [accessed 11 November 2024].

[1016] Legal Ombudsman, 2024. Best practice complaint handling guide. [accessed 11 November 2024].

help keep service providers accountable for processing and determining complaints in a timely manner.

6.195 Third, we consider the ability to opt out of receiving non ephemeral communications regarding complaints will provide complainants with greater agency over their online experience. The opt-out is intended to ensure vulnerable users are not discouraged from reporting due to potential unintended consequences and are empowered to choose whether they receive updates in a way that suits their personal experiences. The opt-out may aid in:

- protecting complainants in a domestic abuse situation who may be subject to harm if the perpetrator intercepts the acknowledgment,
- preventing secondary trauma to children or loved ones if they access updates about a complaint on shared devices,
- preventing re-traumatisation if victims and survivors repeatedly face their trauma through updates about their complaint.

6.196 These benefits contribute to the outcomes this measure seeks to achieve: to encourage user reporting and increase the transparency of reporting and complaints procedures.

## Costs and risks

### Costs

6.197 Service providers will incur costs related to informing complainants that their complaint has been received and providing them with an indicative timeframe for handling the complaint. We expect these costs are likely to be small for most providers.

6.198 We expect that the vast majority of large services and smaller high-risk services will choose to automate this acknowledgement (for example, through an email or pop-up message). We estimated in our November 2023 Consultation that this would require five to 50 days of software engineering time (with an equal amount of time input from professional occupation staff). Using our assumptions on labour costs required for this type of work, we estimate one-off direct costs to be somewhere in the region of £2,000 to £50,000.[1017]

6.199 There would also be some ongoing costs involved in maintaining this measure. Consistent with our standard assumption, we assume annual maintenance costs to be 25% of the initial set-up costs, and estimate these to be approximately £500 to £12,500 per year. We expect service providers with less complex systems and governance processes to incur costs at the lower end of this range. This is likely to include most providers of smaller services.

6.200 Adding functionality to give complainants the choice to opt out of non-ephemeral communications about the complaint could create extra costs for service providers, but we expect the costs to be low. For example, it should be relatively straightforward for providers to offer an opt-out checkbox in both email and webform reporting systems. Providers will also need to amend their systems to ensure that a complainant that has opted out of non-ephemeral communications does not receive any other related communications.

[1017] This is based on our assumptions for labour costs set out in Annex 5. We have updated these estimates since the November 2023 Consultation in line with the latest wage data released by the Office for National Statistics (ONS), as described in that Annex.

### Risks

6.201 While the indicative timeframes set by providers will not be binding, there may be some indirect impacts from communicating these to complainants. For example, a provider may receive repeat complaints if indicative timeframes are not met. While this may incentivise providers to meet their own indicative timeframes, it may also encourage them to set longer timeframes to reduce pressure. We recognise these risks and the pressures they add to a service provider, but consider that the benefits of providing indicative timeframes (ranging from ensuring provider accountability to improving complainant experience) will outweigh them.

6.202 There is a related risk that indicative timeframes may be misunderstood by complainants as a binding deadline (or, as some stakeholders noted, a deadline specific to their complaint). This could result in frustration for complainants if the timeframe is not met. However, we expect the service provider to draft communication about timeframes in a way that does not lead to false expectations.

6.203 There may also be an additional burden of choice on complainants by requiring them to choose whether they receive a response to their complaints. However, we maintain that the potential benefits to vulnerable users (for example, those experiencing domestic abuse) and the reduced risk of harm outweigh this additional burden.

## Rights impact

### Freedom of expression and freedom of association

6.204 We do not consider that this measure would have a negative impact on complainants' rights to freedom of expression or freedom of association.

6.205 We consider it likely that this measure will have a positive impact on the rights of complainants. Transparency and accountability around the complaints process may encourage more complaints, resulting in online spaces becoming safer for users. As the ICO noted, whatever timeframe a service provider sets would need to consider time limits under data protection laws and other areas of law.[1018] This will further protect complainants' rights.

6.206 To the extent that these measures ask a service provider to convey information it might not otherwise convey, there is a potential (but negligible) impact on the service provider's rights to freedom of expression.[1019] We do not consider that this would amount to an infringement of their rights; if it did, we consider this proportionate in the interests of protecting the rights of users and others (including adults and children).

### Privacy

6.207 We do not consider that there are likely to be any additional impacts on complainants' and others' rights to privacy beyond those set out in the first two measures considered in this chapter. This is because we are not recommending that additional personal data is retained or processed to what is needed to handle complaints under other measures.

[1018] ICO response to November 2023 Consultation, pp.18-19.

[1019] We agree with the Centre for Competition Policy's view that providers' rights to freedom of expression should not prevent us from including this measure in the Codes. Centre for Competition Policy response to November 2023 Consultation, p.17.

### Who this measure applies to

6.208 In our November 2023 Consultation, we proposed that these measures should apply to all providers of U2U and search services.

6.209 As with the measure on providing information prior to the submission of a complaint, we are no longer recommending that this measure apply to small, low-risk services. We do not consider it proportionate to apply this measure to such services as the benefits would be small. Our reasoning here is the same for the measure on providing information prior to the submission of a complaint and is set out more fully in paragraphs 6.157-6.159.

6.210 This measure will apply to a provider of a service that is either:

- a large U2U service or a large general search service, or
- at medium or high risk of any kind of illegal harm.[1020]

## Conclusion

6.211 Our analysis shows that this measure will deliver benefits to complainants by increasing their confidence in the service provider's reporting processes and giving them information that may encourage them to engage with complaints processes. The opt-out feature we are recommending will give complainants more choice over their reporting experiences and help avoid unintended consequences from complaints procedures. We also conclude the costs and the impact on rights to be proportionate and limited.

6.212 We have addressed stakeholders' concerns about the challenges of providing and meeting indicative timeframes for different types of complaints. While we acknowledge that a service provider's failure to comply with these timeframes may cause frustration among some complainants, we expect providers to be able to draft their communications in such a way that they do not lead to false expectations, while providing adequate assurance and transparency about how long it may take to process a complaint.

6.213 We are including this measure in our Illegal Content Codes of Practice for U2U and search services and is referred to within these Codes as ICU D4 for U2U services and ICS D3 for search services. The opt-out functionality can be found at ICU D6 for U2U services and ICS D5 for search services. The measure is part of a Codes on terrorism, CSEA and other duties.

## Measure on sending further information about how a complaint will be handled

6.214 In our May 2024 Consultation, we proposed a measure for both the Illegal Content Codes and the Children's Safety Codes, that stated that providers of U2U and search services likely to be accessed by children should include information about the possible outcomes of a complaint, including whether the service provider will update the complainant of the outcome on the complaint, in its acknowledgement of a complaint.[1021]

[1020] This means that this measure will only apply to large vertical search services if they are at medium or high risk of any kind of illegal harm.

[1021] In our May 2024 Consultation, this measure was listed as 'Measure PCU/PCS C4'.

## Summary of stakeholder feedback[1022]

6.215 In response to our May 2024 Consultation, several stakeholders welcomed this measure.[1023]

6.216 Some stakeholders raised concerns about the measure, we have grouped these into the following themes:

- Extent of measure.
- Demand of service providers.

6.217 We detail comments on these themes in the following paragraphs.

### Extent of measure

6.218 We also received feedback about the extent of this measure. Several stakeholders expressed concerns that, if children do not receive an update on the outcome of their complaint, they would lose trust in reporting systems and it could deter them from submitting complaints in future.[1024] We consider this feedback in paragraph 6.223-6.224 in the 'How this measure works' section.

### Demand of service providers

6.219 In contrast, Mid Size Platform Group said this measure could impose disproportionately high resource demands on service providers.[1025] We respond to this feedback in paragraph 6.232-6.233 in the 'Costs and risks' section and paragraph 6.236 in the 'Who this measure applies to' section.

## Our decision

6.220 We have decided to broadly confirm the measure we proposed in the May 2024 Consultation, but have made some clarificatory amendments and some changes:

- For U2U services, the measure will apply to a provider of a service likely to be accessed by children that is either large, or at medium or high risk for any kind of illegal harm.
- For search services, the measure will apply to a provider of a service likely to be accessed by children that is either a large general search service, or at medium or high risk of any kind of illegal harm. As discussed in paragraph 6.187, we have added a provision (ICU D6 for U2U services and ICS D5 for search services) recommending that service providers allow users to opt out of receiving communications from the provider

[1022] Note: this list is not exhaustive and further responses can be found in Annex 1.

[1023] Commissioner Designate for Victims of Crime Northern Ireland response to May 2024 Consultation, p.6; CELCIS response to May 2024 Consultation, p.15; Dean, J. response to May 2024 Consultation, pp.16-17; Kooth Digital Health response to May 2024 Consultation on Protecting Children from Harms Online, p.12; NSPCC response to May 2024 Consultation, p.58.

[1024] Children's Commissioner for Wales response to May 2024 Consultation on Protecting Children from Harms Online, p.3; Dean, J. response to May 2024 Consultation, pp.16-17; NSPCC response to May 2024 Consultation, p.58; Scottish Government response to May 2024 Consultation on Protecting Children from Harms Online, p.16; The Northern Ireland Commissioner for Children and Young People (NICCY) response to May 2024 Consultation on Protecting Children from Harms Online, p.34. We note the NSPCC made a similar call for services to be required to update users on the outcome of their reports and complaints in response to the November 2023 Consultation. NSPCC response to November 2023 Consultation, p.31.

[1025] Mid Size Platform Group response to May 2024 Consultation, p.11.

after they have submitted a complaint. As this measure relates to communication related to a complaint, the opt-out provision also applies here.

6.221 The full text of the measure can be found in our Illegal Content Codes of Practice for U2U and search services, and is referred to within these Codes as ICU D5 for U2U services and ICS D4 for search services. This measure is part of our Codes on terrorism, CSEA and other duties.

# Our reasoning

## How this measure works

6.222 When acknowledging complaints, providers of U2U and search services likely to be accessed by children should provide the complainant with an explanation of what actions may be taken in response to the complainant's complaint and if, and when, the complainant should expect to hear the outcome. This information need only explain the provider's general position on responding to complaints and does not need to be personalised.

6.223 This will improve the transparency of a provider's complaints procedure and could help reassure complainants, particularly children, that providers are handling their complaints.

6.224 We consider service providers to be best placed to decide the most effective format and wording to include in this information. To achieve the aim of improving transparency and increasing complainants' trust in complaints mechanisms (in line with our measure that recommends complaints systems to be easy to access and use), providers should ensure the information included in their acknowledgement of complaints is comprehensible and accessible to all users (including children), taking into consideration other accessibility requirements identified in their risk assessment.

6.225 We are not suggesting that providers must update complainants about the outcome of their complaint. This measure leaves that to the discretion of a provider, as we recognise that some may be better resourced to do that than others. We maintain that notifying complainants of *whether* they will hear about the outcome of their complaint is sufficient to meet transparency duties in the Act. This will help users to understand what to expect from a complaints procedure.

6.226 Service providers in scope of this measure should allow complainants to opt out of communications relating to complaints.[1026] As such, complainants should be able to choose whether they not to receive information about the outcome of a complaint, if the service shares this information.

6.227 As described in paragraph 6.187, providers have flexibility over how to implement this opt-out. Providers do not need to provide an opt-out option for acknowledgements that disappear once the user has viewed them and cannot be restored.

## Benefits and effectiveness

6.228 As set out in our May 2024 Consultation, our research (referred to in paragraph 6.150) suggests that when children are not informed of the outcome of their complaints, they are discouraged from complaining again.[1027] This is because the lack of response causes

[1026] For U2U services, this is set out in our Codes as measure ICU D6. For search services, this is ICS D5.

[1027] Ofcom, 2024. Children's attitudes to reporting content online, p.39. [accessed 11 November 2024].

children to believe no action has been taken. Participants in our 2024 research into children's attitudes to reporting said services should update the user on the progress of their report and next steps, including when they should expect to receive a response.[1028] More generally, informing users of the action taken – or the actions that might be taken – as a result of a complaint could make reporting and complaints procedures easier to use for children.[1029]

6.229 By including information about how complaints are handled in their acknowledgement of complaints, providers can dispel the misconception that a lack of subsequent communication means no action is being taken. It could also encourage complainants and prospective complainants to engage in the reporting and complaints process again in the future.

6.230 This measure seeks to improve confidence in complaints procedures by increasing transparency about, and understanding of, the possible outcomes from a complaint and what a complainant should expect from a complaints procedure. We expect this to encourage children who have encountered illegal or otherwise harmful content to complain, thereby increasing the proportion of such content which is removed.

6.231 Communicating the potential outcomes and actions a service provider might take could also help educate users on what content is illegal and what is not, which may improve the quality of complaints over time.

## Costs and risks

6.232 As set out in our May 2024 Consultation, we expect that providers may incur a small incremental cost as a result of this measure. These costs would be in addition to the costs outlined in the measure relating to acknowledging complaints and providing indicative timeframes.[1030] The incremental cost of this measure would result from including in a complaint acknowledgement an explanation of what actions may be taken in response to the complaint and when the complainant should expect to hear the outcome. As this information would not need to be personalised to a given complaint or complainant, we expect these costs would be small.

6.233 We expect providers to incur costs in agreeing the actions to take in response to complaints and in getting these signed off through their internal governance processes. We also expect providers to incur a cost in drafting an explanation of these actions for inclusion in the acknowledgement of complaints. We expect both these costs to be negligible.

## Rights impact

### Freedom of expression, freedom of association and privacy

6.234 We do not consider that this measure would have any adverse impacts on complainants' rights to freedom of expression, freedom of association or privacy. The impact of recommending the provider to convey information it might not otherwise convey is negligible.

[1028] Ofcom, 2024. Children's attitudes to reporting content online, p.39. [accessed 11 November 2024].

[1029] Department for Science, Innovation and Technology (DSIT) and Department for Culture, Media and Sport (DCMS), 2021. Child online safety: Age-appropriate content, [accessed 10 November 2024].

[1030] Ofcom, 2024. Protecting Children from Harms Online, Volume 5: What should services do to mitigate the risks of online harms to children?, p.257. [accessed 11 November 2024].

### Who this measure applies to

6.235 In our May 2024 Consultation, we proposed that this measure should apply to all providers of U2U and search services likely to be accessed by children.

6.236 As with the measures on providing information prior to the submission of a complaint and on sending acknowledgements and indicative timeframes, we are no longer recommending that this measure apply to smaller, low-risk services. Our reasoning is the same as for the previous measures and is set out more fully in paragraphs 6.157-6.161.

6.237 For services likely to be accessed by children, this measure will apply to a provider of a service that is either:

- a large U2U service, or large general search service; or
- at medium or high risk of any kind of illegal harm.[1031]

## Conclusion

6.238 Our analysis shows that this measure will deliver benefits by strengthening reporting and complaints procedures for both service providers and complainants. It will reassure complainants that their complaints are being taken seriously and will encourage service providers to operate transparent and accountable complaints procedures. As explained, we expect that the marginal cost of this measure will be limited. On balance, we therefore consider it to be proportionate.

6.239 We have amended this measure so that it does not apply to small, low risk services. The measure will only apply to U2U services likely to be accessed by children that are large, or at medium or high risk of any kind of illegal harm, and search services likely to be accessed by children that are a large general search service, or at medium or high risk of any kind of illegal harm. We consider the costs to be proportionate to these providers.

6.240 We are including this measure in our Illegal Content Codes of Practice for U2U and search services, and is referred to as ICU D5 for U2U services and ICS D4 for search services. It is part of our Codes on terrorism, CSEA and other duties.[1032]

# Measures on appropriate action for processing relevant complaints

6.241 These next two sections discuss measures that consider the appropriate action services should take to process complaints.

6.242 In this section we discuss measures on appropriate action for complaints about: suspected illegal content; the use of proactive technology; and a third category of complaints we refer to as “all other relevant complaints”. This third category comprises complaints regarding providers’ non-compliance with the duties in the Act: the safety duty, content reporting, freedom of expression and privacy.

[1031] This means that this measure will only apply to large vertical search services if they are at medium or high risk of any kind of illegal harm.

[1032] Complainants will be able opt out of such communications (ICU D6/ICS D5).

6.243 Sections 21 and 32 of the Act require all regulated U2U and search services to operate complaints procedures that ensure appropriate action is taken in response to reports about illegal content and other types of relevant complaints. The nature of the appropriate action will depend on the type of complaint.

6.244 In our November 2023 Consultation, we proposed measures relating to appropriate action for processing relevant complaints. These measures included:

- a measure recommending that U2U and search services handle relevant complaints about suspected illegal content in accordance with its prioritisation process and performance targets (or promptly if these do not exist), and act in accordance with our recommendations on content or search moderation functions.
- a measure recommending that U2U and search services should (where relevant) inform complainants of their right to bring proceedings if the user believed that the use of proactive technology had resulted in: content being taken down, given a lower priority or access to it being restricted; search content being deindexed or downranked; or if the technology had been used in a way that is in breach of a service provider's terms of service or publicly available statements.
- a measure recommending that U2U and search services should establish a triage process for relevant complaints.

6.245 We received cross-cutting feedback for these measures and so address them together in this section.

## Summary of stakeholder feedback[1033]

6.246 Some stakeholders expressed support for the measures discussed in this section.[1034] We have grouped other stakeholder responses into the themes below:

- Taking appropriate action in response to spam complaints or complaints not about illegal content.
- Requests for more detailed guidance on how to handle complaints.

6.247 We detail comments on these themes in the paragraphs below.

### Taking appropriate action in response to spam complaints

6.248 In response to the measure on acknowledgements and indicative timeframes, Meta highlighted that it receives high volumes of unfounded or meritless complaints which can "act to the detriment of those genuinely reporting harmful or illegal content" and suggested that our measures include an exception for reports identified as 'spam'.[1035] Google also suggested that our measures could be exploited by perpetrators who might seek to obtain information about how to circumvent detection systems. Google suggested that our measures should be amended so that they do not apply to spam complaints or

[1033] Note: this list is not exhaustive and further responses can be found in Annex 1.

[1034] Canadian Centre for Protection of Children response to November 2023 Consultation, p.21; Federation of Small Businesses response to November 2023 Consultation, p.3; Global Partners Digital response to November 2023 Consultation, pp.13-18; Refuge response to November 2023 Consultation, pp.11-12; Snap response to November 2023 Consultation, p.13; Welsh Government response to November 2023 Consultation, p.4.

[1035] Meta response to November 2023 Consultation, p.28.

complaints by providers of malware.[1036] We consider these concerns in paragraph 6.267-6.280 in the 'How these measures work' section.

6.249 Global Partners Digital also called for "appropriate safeguards and verification measures" to ensure that complaints procedures are not abused or misused by malicious actors seeking, for example, to censor content.[1037] One individual respondent argued that reporting and complaints procedures can be misused by "malicious actors" who may use complaints to "silenc[e] marginalised communities" such as sex workers.[1038] Another respondent [✂] noted that services that host adult content often receive reports and complaints that seek to undermine sex workers (or the adult industry) rather than legitimately highlight illegal content.[1039] We address this feedback in paragraphs 6.267-6.280 in the 'How this measure works' section and paragraph 6.283 in the 'Benefits and effectiveness' section.

### Requests for more detailed guidance on how to handle complaints

6.250 This theme was most prominent in responses from civil society stakeholders, who called for minimum standards or further information on the implementation of 'appropriate action' and how to handle complaints.[1040] Global Partners Digital provided suggestions for specific steps providers should take to improve their complaints processes (including their handling of appeals).[1041] We consider this feedback in paragraph 6.256 in the section 'How these measures work'.

## Our decision

6.251 We have decided to broadly confirm the measures we proposed in the November 2023 Consultation, with some changes in response to stakeholder feedback and clarificatory amendments:

- We have made some clarificatory amendments to our measure on appropriate action for relevant complaints about suspected illegal content that set out that, when a provider receives a relevant complaint about suspected illegal content, it should treat the complaint as reason to suspect that the content may be illegal and review it in accordance with the relevant content and search moderation measures. We have amended the drafting of this measure so it does not unnecessarily repeat content moderation measures on prioritisation and performance targets, which apply anyway. However, for providers which are not subject to those recommendations, we continue to recommend that they should consider the complaint promptly.[1042]
- We have amended our measure on appropriate action for relevant complaints about proactive technology. For both U2U and search services, the measure says that a provider should inform the complainant of the action the provider may take in response to complaints about the use of proactive technology. For U2U services, the measure also says the provider should inform the complainant of their right to bring proceedings. For search services, we also clarify when or how a search service may have used

[1036] Google response to November 2023 Consultation, pp.54-55.

[1037] Global Partners Digital response to November 2023 Consultation, p.16.

[1038] Are, C. response to November 2023 Illegal Harms Consultation, pp.12-13.

[1039] [✂].

[1040] 5Rights Foundation response to November 2023 Consultation, pp.23-24; Global Partners Digital response to November 2023 Consultation, pp.16-18; Refuge response to November 2023 Consultation, pp.11-12.

[1041] Global Partners Digital response to November 2023 Consultation, pp.16-18.

[1042] In our Codes, this measure is ICU D7 for U2U services and ICS D6 for search services.

proactive technology in a way not contemplated by or in breach of a publicly available statement.[1043]

- We have amended our measure on appropriate action for all other relevant complaints to make clear what types of complaints the measure applies to and how a service provider should deal with those complaints.[1044]
- Reflecting on stakeholder feedback on our measures about appropriate action, we have decided to allow a service provider to disregard a relevant complaint (excluding a complaint that is an appeal) if it determines it to be manifestly unfounded, should it have a policy in place to do so.[1045]

6.252 These measures are part of our Illegal Content Codes of Practice for U2U and search services. The full text of the measures can be found in our Codes:

- Our measure on appropriate action for relevant complaints about illegal content is referred to as ICU D7 for U2U services and ICS D6 for search services.
- Our measure on appropriate action for relevant complaints about the use of proactive technology is referred to as ICU D11 for U2U services and ICS D10 for search services.
- Our measure on appropriate action for all other relevant complaints is referred to as ICU D12 for U2U services and ICS D11 for search services.
- Our exception for relevant complaints (excluding appeals) that a provider determines to be manifestly unfounded is referred to as ICU D13 for U2U services and ICS D12 for search services.

6.253 These measures form part of our Codes on terrorism, CSEA and other duties.

# Our reasoning

## How these measures work

6.254 The measures in this section outline the appropriate action providers should take in response to complaints. The nature of this action varies depending on the type of complaint made.

6.255 For these measures, the complaint types are:

- complaints about suspected illegal content,
- complaints about the use of proactive technology in breach of the provider’s terms or policies, and
- all other relevant complaints:
  > Complaints about non-compliance with the illegal content safety duty or content reporting duty; and
  > Complaints about non-compliance with freedom of expression or privacy duties.

6.256 As set out in paragraph 6.250, several stakeholders requested more detailed guidance on how to handle complaints. In chapter 2 of this Volume: ‘Content moderation’, we explain that we are not in a position to do so at this early stage in the regulatory regime. We

[1043] In our Codes, this measure is ICU D11 for U2U services and ICS D10 for search services.

[1044] In our Codes, this measure is ICU D12 for U2U services and ICS D11 for search services.

[1045] In our Codes, this measure is ICU D13 for U2U services and ICS D12 for search services.

consider it appropriate for our first iteration of the Codes to explain the content reporting duty in the Act and establish a baseline on which we can build in future, if necessary, after monitoring implementation of these measures.

6.257 We now set out how each of the measures in this section work.

### Measure on appropriate action for relevant complaints about suspected illegal content (ICU D7/ICS D6)

6.258 This measure ensures relevant complaints about illegal content on U2U and search services are handled in accordance with relevant moderation measures.

6.259 One of the ways we expect a provider to identify suspected illegal content is through reporting and complaints. Therefore, complaints about illegal content should be handled as suspected illegal content as per measure ICU C1/ICS C1, unless the complaint is manifestly unfounded. This means that the appropriate action for a complaint about suspected illegal content is usually to action it as per the moderation measures in chapters 2 and 3 of this Volume: 'Content moderation' and 'Search moderation'.

6.260 As per the moderation measures, providers of large and/or multi-risk services should also establish and apply performance targets and prioritisation processes. However, some providers are not in scope of these content or search moderation measures. For them, we have stipulated in this complaints measure that complaints should be reviewed promptly. This will encourage all service providers to consider the resolution of complaints a priority.

### Measure on appropriate action for relevant complaints about proactive technology (ICU D11/ICS D10)

6.261 The Act requires that providers take appropriate action in response to certain complaints about the use of proactive technology:

- For a U2U service, these complaints can be made by a complainant where the use of proactive technology means their content has been taken down or other users' access to it restricted, or it becomes less visible, and the user considers that the technology has been used in a way not contemplated by, or in breach of, the terms of service.
- For a search service, these complaints can be made by an interested person where the use of proactive technology means content relating to them no longer appears in search results or is given a lower priority in search results and the interested person considers that the technology has been used in a way not contemplated by, or in breach of, the provider's policies.

6.262 The basis of a complaint about the use of proactive technology is therefore not necessarily about the nature of the content taken down or the use of proactive technology per se, but whether the operation of the proactive technology concerned is consistent with the terms of service or publicly available statement.

6.263 As it relates to search services, we have amended this measure to remove the recommendation that providers of search services inform interested persons where relevant of their right to bring legal proceedings. Interested persons who are not users of the service do not typically have a contract with the service provider and so do not have a right to bring proceedings for breach of contract.

6.264 Instead, we have added a recommendation that the appropriate action for both U2U and search services would be for a provider to inform complainants of the action (if any) the provider may take in response to their complaint. This does not need to be personalised to the complainant or involve a direct communication with the complainant.

### Measure on appropriate action for all other relevant complaints (ICU D12/ICS D11)

6.265 This measure ensures all other relevant complaints (apart from appeals, which are considered in the section below), are directed towards the most relevant individual or team to be processed. It recommends a provider nominate an individual or team who is responsible for securing this.

6.266 It also recommends that such complaints be dealt with in a way that protects users and within timeframes the provider has determined are appropriate. As per our measure on sending acknowledgements and providing indicative timeframes for handling complaints, the provider should determine what timeframe would be appropriate for considering the complaint.

### Exception for "manifestly unfounded" complaints (ICU D13/ICS D12)

6.267 As summarised in paragraph 6.248 we received feedback from stakeholders about 'spam' complaints and the impact it can have on the service's ability to handle complaints. We acknowledge that providers can receive a large volume of complaints about illegal content, not all of which necessarily reflect a genuine or well-founded concern held by the complainant. We have also considered whether providers may receive other types of relevant complaints which are spam complaints.

6.268 For example, there is evidence that unfounded complaints can arise from coordinated (malicious) mass reporting. In its Quarterly Adversarial Threat Report, Meta highlighted the extent of 'mass reporting' – which it describes as coordinated reporting intended to abuse their reporting systems to get accounts or content incorrectly taken down – in several countries (and across several services), usually for political purposes.[1046] The Oxford Internet Institute also highlighted that mass-reporting of content or accounts can be used to 'censor speech and expression'.[1047]

6.269 Mass reporting can also be used to target service providers. For instance, a new or small service provider could be targeted by a competitor, by supporters of a competitor's platform, or by malicious actors to overwhelm the provider's moderation systems or administration. One stakeholder referenced this concern in response to our November 2023 Consultation.[1048] In such instances, this may prevent the provider from actioning relevant complaints about illegal content.

6.270 We accept that it is not necessarily appropriate for a provider to consider all complaints, especially if it is likely to receive a significant number of manifestly unfounded complaints, or where it is in a position to determine that a complaint is manifestly unfounded. Doing so could cause harm to users as a provider's resources may be diverted away from considering content that causes serious harm. As evidenced above, such complaints could also pose a risk to users' rights to freedom of assembly and expression.

6.271 We expect services to handle all relevant complaints as outlined in the Act. This is required by the Act and is important for keeping users safe online. However, we recognise that the

[1046] Meta. 2023. Quarterly Adversarial Threat Report [accessed 6 November 2024].

[1047] Bradshaw, S., Bailey, H., and Howard, P. 2021. Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation, p.17 [accessed 27 November 2024].

[1048] [✂].

requirement to enable complaints and handle them can be resource-intensive, especially if the complaints are unfounded or 'spam', or if the complaints procedure is abused by malicious actors.

6.272 The measure we proposed in our November 2023 Consultation did not give providers the option to disregard any complaints. Recognising that services may receive complaints about illegal content which may be manifestly unfounded, we are including within our Codes a provision that provides guidance on how a service provider can handle complaints about relevant complaints that it has determined to be manifestly unfounded.

6.273 We refer to 'manifestly unfounded' complaints rather than 'spam' complaints because we consider the term "spam" could be confusing: it may be misread as referring to complaints *about* spam, for example. We consider that the term 'manifestly unfounded' sets an appropriately high bar for the provider to meet before it can decide to disregard a relevant complaint. A complaint is not 'manifestly unfounded' merely because it is incorrect. It is manifestly unfounded when it is possible to infer that the person is engaging in coordinated (malicious) mass reporting t, or when the complaint is clearly not a relevant type of complaint.

6.274 Complaints which are appeals are excluded from this amendment. It is important that a service provider is not able to disregard complaints which are appeals because these are safeguards for freedom of expression, and the only recourse available to users whose access to the service may have been restricted. By definition, an appeal can only be submitted by a complainant where action has already been taken against them because of a decision that content is illegal. We therefore do not consider that the same concerns about mass reporting arise in relation to appeals.

6.275 The amendment also does not apply to complaints from trusted flaggers. It would be hard for malicious actors to abuse or misuse a channel dedicated to trusted flaggers since they should not have access to it.

6.276 A service provider does not need to disregard manifestly unfounded complaints. If a provider chooses to do so, it should only disregard complaints if:

a) it has a policy setting out attributes and information it requires to make a decision that a complaint is manifestly unfounded;
b) the complaint is determined in accordance with that policy; and
c) it regularly reviews and, where appropriate, makes changes to its policy to ensure the accuracy of decision-making.

6.277 Our measure allows services to determine the attributes and information it will use to make such a decision on a complaint. While we are not prescriptive about what attributes or information a service provider includes in its policy, the term 'manifestly unfounded' means that the threshold must be high.

6.278 If a provider chooses to develop and implement a policy for manifestly unfounded complaints, we recommend that it reviews the outcomes of this policy annually. This is consistent with our recommendation for how often risk assessments are reviewed. We recognise that regular reviews could present challenges for small services. However, we consider that regularly reviewing the policy is necessary to ensure that legitimate relevant complaints are not being disregarded.

6.279 Service providers should not implement this exception in a way that results in complaints which may be well-founded being disregarded. We expect the high threshold we are setting for 'manifestly unfounded' complaints to aid providers in ensuring this. This will also act as a safeguard to users. We have set out that a policy for manifestly unfounded complaints should also include a regular review process for similar reasons. This review process should allow a service to assess the accuracy of the decisions made as a result of the policy and where necessary update the policy so that the chances of relevant complaints being identified as 'manifestly unfounded' incorrectly remains low.

6.280 We consider that this exception will enable a provider to take appropriate action on complaints and to filter out complaints which are manifestly unfounded, such as those from coordinated malicious actors.

## Benefits and effectiveness

6.281 Our measures on taking appropriate action for relevant complaints about suspected illegal content are designed so a service provider can swiftly determine and take action in response to complaints and where appropriate, subsequently remove illegal content from a service in a timely and efficient manner.[1049] By setting out how complaints procedures feed into content or search moderation functions, and recommending services ensure that complaints are dealt with by appropriate individuals or teams, the process of determining and actioning complaints is made more efficient. This will reduce the potential of illegal content reaching further users and will also reassure complainants that the service provider takes swift action in regard to complaints.

6.282 These measures will also aid the effectiveness of complaints procedures by outlining a clear process for how complaints should be dealt with. This should help ensure complaints are responded to within appropriate timeframes.

6.283 For our exception for manifestly unfounded complaints, we are setting a high threshold for providers to decide they do not need to consider a complaint.[1050] As explained in paragraph 6.279, we intend this high threshold to act as a safeguard to complainants' rights. We expect the service provider to be able to explain and justify the basis on which it determines complaints to be manifestly unfounded.

6.284 The measure on appropriate action for relevant complaints about proactive technology will ensure transparency over how a provider treats these complaints and make it clear to users of U2U services what their rights are.[1051]

6.285 Overall, these measures will lead to more effective complaints procedures and better outcomes for users.

## Costs and risks

### Costs

6.286 Reporting and complaints measures apply to all service providers within scope of the Act. By avoiding overly specific recommendations, we have aimed to make our recommendations both proportionate to, and suitable for, a wide range of services.

[1049] In our Codes, these measures are ICU D7 for U2U services and ICS D6 for search services; and ICU D12 for U2U services and ICS D11 for search services.

[1050] In our Codes, the amendment for manifestly unfounded complaints can be found in ICU D13 for U2U services and ICS D12 for search services.

[1051] ICU D11 for U2U services and ICS D10 for search services.

6.287 As explained in the November 2023 Consultation, we expect the costs of implementing these measures to vary depending on service type, size, and risk level. While costs may be significant for some service providers, we consider them to be largely the result of requirements of the Act over which we have little discretion. In considering our recommendations, we have taken into account the likely cost burden to services and have not been prescriptive in setting specific timescales for handling complaints, allowing providers the flexibility to determine what is appropriate for their service.

6.288 We expect the potential volume of complaints about illegal content to vary with the size of the service (while acknowledging that the type of service may also be a factor). This means providers incurring the highest costs are likely to be those with the greatest ability to absorb them. Complaints about illegal content are likely to vary with the volume of content being shared by users (for U2U services) and with the volume of search queries (for search services). Where a service provider receives a large volume of complaints relating to illegal content, we consider that these costs could be regarded as part of content or search moderation. While these costs may be significant for some services, a service provider would not be able to comply with the duty in section 10(3)(b) of the Act if it did not consider the complaints.

6.289 We acknowledge that making illegal content judgments creates potential added complexity for a service provider. The Act does not require providers to make illegal content judgments if they are satisfied that their terms of service or community guidelines prohibit all content that would be considered illegal in the UK. If a service provider chooses to use its terms of service to make decisions in this way, this would reduce the cost burden compared to a situation where the provider had to make formal illegal content judgements for every decision.

6.290 As outlined in paragraphs 6.267-6.280, we have included in our Codes an exception that allows service providers to disregard complaints they have determined to be manifestly unfounded. Although there may be costs associated with establishing a policy and process for identifying such complaints, and for reviewing this policy, we expect that providers who choose to do so would mitigate some of the costs incurred by our measures on handling relevant complaints about illegal content and on content or search moderation (ICU C1-2, ICU D7/ICS C1, ICS D6) as they will likely be processing fewer complaints. Because such a policy is both optional and may result in cost savings for providers, we have not explored the potential costs in detail.

### Risks

6.291 There is a risk to complainants that some complaints may be wrongly regarded as manifestly unfounded and not dealt with. We are balancing this risk against the risk that a service provider is over-burdened by having to put manifestly unfounded complaints through its content or search moderation processes (and the knock-on impact this could have on the provider effectively handling other complaints). We expect the high threshold we are establishing to protect against this risk to complainants. As explained in the 'How these measures work' section, we expect providers to consider these potential risks when establishing their policies for identifying manifestly unfounded complaints.

## Rights impact

### Freedom of expression and freedom of association

6.292 We do not consider our measures would have any negative impact on the rights of users, affected persons, interested persons or services to freedom of expression. The impacts of the decisions service providers take following complaints about suspected illegal content are considered as a part of our assessment of our measures on content or search moderation, especially measures ICU C1, ICS C1 and ICU C2.

6.293 These measures will help give complainants confidence that appropriate action will be taken in response to their complaints, and help to keep users safe online by facilitating the removal of illegal content. Therefore, they are likely to have a positive impact on the rights to freedom of expression and of association, and to other human rights which may be engaged by illegal content.

6.294 We recognise that there is some risk that providers may wrongly identify complaints as manifestly unfounded, and that the harms complainants may experience as a result may engage their human rights. However, we are considering this risk against the evidenced risk that complaints processes will be used maliciously. This could be against particular users, for example to inhibit their freedom of expression (including, in particular, political expression) or lawful commercial activities. Or it could be done to cause harm to a service provider by raising its costs and deployment of resources. As a consequence, well-founded complaints might not be considered, which would expose users to harm; or the service provider may leave the market, depriving users of a platform on which to express themselves and associate with one another. Having weighed up the various impacts and rights, we consider that the impact of our approach is proportionate.

### Privacy

6.295 We do not consider that these measures would give rise to any additional impacts on complainants' and others' rights to privacy beyond those already set out in the other measures considered in this chapter. These measures could have a positive impact on individuals' right to privacy by providing greater transparency and accountability around decisions made in respect of the illegal content safety duties.

6.296 In implementing these measures, service providers should ensure they comply with data protection laws and familiarise themselves with any relevant guidance issued by the ICO.[1052] We consider that service providers can and should implement the measures in a way which minimises the amount of personal data that is processed (in line with the principle of data minimisation, which our amendment for manifestly unfounded complaints supports).[1053]

## Who these measures apply to

6.297 Reporting and complaints duties apply to all service providers within the scope of the Act, and all in-scope services are required to take appropriate action in response to relevant complaints. Our measures on appropriate action for complaints considered in this section apply to all service providers.

[1052] Information Commissioner's Office, UK GDPR guidance and resources, [accessed 10 November 2024].

[1053] Information Commissioner's Office, 2023. A guide to the data protection principles, [accessed 10 November 2024].

## Conclusion

6.298 The measures in this section set out expectations for a service provider's policies on taking appropriate action in response to complaints. These measures are an important component of services' moderation functions and will contribute to ensuring that providers comply with their other duties under the Act, and help safeguard users' rights in relation to the use of proactive technology. Ultimately, they will play an important role in keeping users safe online. Whilst the measures could in some cases have significant costs, we consider these are necessary to comply with providers' duties under the Act.

6.299 Having assessed stakeholder responses, we have added an exception to allow service providers to disregard relevant complaints (excluding appeals) that it has determined to be manifestly unfounded provided that they have a policy in place to do this, and that they review this policy periodically. We consider that this amendment will minimise burdens on service providers and avoid the risks to users which may come from malicious use of complaints procedures. However, it does not allow service providers to avoid meeting their obligations to take appropriate action in response to complaints.

6.300 We are including these measures in our Illegal Content Codes of Practice for U2U and search services.

- Our measure on appropriate action for relevant complaints about suspected illegal content is referred to as ICU D7 for U2U services and ICS D6 for search services.
- Our measure on appropriate action for relevant complaint about proactive technology (which are not appeals) is referred to as ICU D11 for U2U services and ICS D10 for search services, and
- Our measure on appropriate action for all other relevant complaints (except appeals) is referred to as ICU D12 for U2U services and ICS D11 for search services.
- Our measure setting out the exception for relevant complaints (except appeals) that allows providers to disregard a complaint it has determined to be manifestly unfounded is referred to as ICU D13 for U2U services and ICS D12 for search services.

6.301 These measures are part of our Illegal Content Codes on terrorism, CSEA and other duties.

# Measures on appropriate action for relevant complaints which are appeals

6.302 In this section we discuss three measures about appropriate action for relevant complaints which are appeals. Appeals are relevant complaints which can be made when a provider takes action against content, search content or a user because it considers that content is illegal content.[1054]

6.303 In our November 2023 Consultation, we proposed two measures relating to the determination of appeals and one measure about action following determination.

6.304 On measures relating to determination of appeals, one of these measures applied to large or multi-risk services, and the other applied to all other services. In our May 2024

[1054] Section 21(4)(c), section 21(4)(d), section 32(4)(c) and section 32(4)(d) of the Act.

Consultation, we made some changes to these measures. As such, the two proposed measures about determination of appeals recommended:

- A provider which is large[1055] or multi-risk should set and monitor its performance for determining relevant complaints which are appeals against performance targets. These targets should relate to at least the time it takes to determine the appeal and the accuracy of decision making. When determining what priority to give to its review of an appeal, the provider should consider:
  > the seriousness of the action taken against the user or content;
  > whether the initial decision was made by content identification technology; and
  > the past error rate for illegal content judgements.
- Providers which are neither large nor multi-risk should determine appeals promptly.

6.305 On our measure regarding action following appeals, in our November 2023 Consultation we proposed this to all providers and search services.

6.306 In our May 2024 Consultation, we also made some changes to our measure about action following the determination of appeals and proposed:

- If a provider reverses a decision that content or search content was illegal content, it should reverse the action taken against the user or the content (or interested person or search content) so far as appropriate. It should also adjust the relevant moderation guidance where necessary to avoid similar errors in future, and take steps within its power to secure that the use of automated content moderation technology does not cause the same content to be taken down again or the same search content to no longer appear in search results or be given a lower priority.

6.307 Our proposed recommendations were broadly the same for both U2U and search services. Our proposals reflected the fact that the types of person who can bring appeals under the Act are different for U2U services (users) and search services (interested persons). The types of action a provider can take in response are also slightly different for U2U services and search services.

6.308 We received feedback that was relevant to each of these measures and therefore address them together in the following sections.

## Summary of stakeholder feedback[1056]

6.309 Stakeholders generally supported that our Codes included measures about appeals, with most of those who commented requesting clarifications or further guidance on how the measures should be implemented.[1057] We have grouped responses by themes:

- Independent appeals process.

[1055] For search services this refers to large general search services.

[1056] Note: this list is not exhaustive and further responses can be found in Annex 1.

[1057] Big Brother Watch response to November 2023, p.9; Name withheld 5 response to November 2023 Illegal Harms Consultation, p.15; Federation of Small Businesses response to November 2023 Consultation, p.3; Glitch response to November 2023 Consultation, p.10; Global Partners Digital response to November 2023 Consultation, p.16; Google response to November 2023 Consultation, pp.49-52; Online Dating and Discover Association response to November 2023 Consultation, p.2; Open Rights Group response to November 2023 Consultation, p.4; Snap response to November 2023 Consultation, pp.13-14; Ukie response to November 2023 Consultation, p.22; WeProtect Global Alliance response to November 2023 Consultation, pp.23-25.

- The obligation to accept appeals, including on downranked search content and video-sharing platforms (VSPs).
- Recommended criteria providers should consider when prioritising appeals.
- Actions following determination of appeals.

6.310 We detail comments on these themes in the paragraphs below.

### Independent appeals process

6.311 Many civil society stakeholders called for us to play a role in service providers' complaints processes, suggesting Ofcom should handle complaints from users or offer an independent appeals process.[1058] Other stakeholders called for us to establish an alternative dispute resolution procedure for service providers.[1059] We consider this feedback in paragraphs 6.322-6.323 in the 'How these measures work' section.

### The obligation to accept appeals, including on downranked search content and video-sharing platforms (VSPs)

6.312 Google raised concerns about the obligation to accept appeals about downranked search content and non-video content on VSPs.[1060] It requested that appeals about search content should be limited to illegal content and "delisted" search content, rather than downranked search content. We have considered this feedback in paragraph 6.324 in the 'How these measures work' section.

### Recommended criteria providers should consider when prioritising appeals

6.313 Snap agreed that appeals should be prioritised and suggested additional criteria that might be considered.[1061] Snap also highlighted that this might create operational challenges that would require regular updates. We consider this feedback in paragraph 6.329 in the 'How these measures work' section.

### Actions following determination appeals

6.314 While they largely agreed with our recommendations on actions that should follow the reversal of a content moderation decision, Snap raised concerns about how the actions would be implemented across services where reversal and adjustment of content moderation guidance might not be possible.[1062] We consider this feedback in paragraph 6.331 in the 'How these measures work' section.

## Our decision

6.315 We have decided to proceed with these measures broadly as proposed in our May 2024 Consultation with some amendments to two of the measures.

[1058] 5Rights Foundation response to November 2023 Consultation, p.15; Bereaved Families for Online Safety response to November 2023 Consultation, p.1; Cybersafe Scotland response to November 2023 Illegal Harms Consultation, p.10; Name withheld – a civil society organisation response to November 2023 Illegal Harms Consultation, p.3; New Zealand Classification Office response to November 2023 Consultation, p.9; The Cyber Helpline response to November 2023 Consultation, p.11.

[1059] SWGfL response to November 2023 Consultation, p.15; UKSIC response to November 2023 Consultation, pp.49-50, 54-55.

[1060] Google response to November 2023 Consultation, pp.50-51.

[1061] Snap response to November 2023 Consultation, p.14.

[1062] Snap response to November 2023 Consultation, pp.14-.15.

6.316 We have amended our measure for providers of large and/or multi-risk services which sets out appropriate action on determination of appeals to clarify that a service provider should have in place a policy that sets out how appeals will be prioritised, as opposed to considering the factors we recommend when handling each individual appeal. In setting the policy:

- Providers of U2U services that are large and/or multi-risk should consider the seriousness of the action taken against the user or in relation to the content, whether content identification technology was used in decision making, and the past error rate on the service in relation to illegal content judgements.
- Providers of search services that are large general search services and/or multi-risk should consider the seriousness of the action taken in relation to the search content, whether content identification technology was used in decision making, and the past error rate on the service in relation to illegal content judgements.[1063]

6.317 We have amended our measure on action following determination of appeals to clarify that:

- For U2U services, if a provider reverses a decision that content was illegal content, it should: so far as appropriate and possible reverse the action taken against the user or content for the purpose of restoring the position of the content or user to what it would have been; adjust any relevant content moderation guidance if appropriate to ensure it is accurate where there is significant evidence of content being taken down in error; and, where possible and appropriate, take steps to secure that the use of automated content moderation technology does not result in the same content being taken down again.
- For search services, if a provider reverses a decision that search content was illegal content it should: so far as appropriate and possible, reverse the action taken in relation to the search content for the purpose of restoring its position to what it would have been had the decision not been made; adjust any relevant search moderation guidance if appropriate to ensure it is accurate where there is significant evidence of search content not appearing in search results of being given a lower priority in the ranking of search results; and where possible and appropriate, take steps to secure that the use of automated content moderation technology does not result in the same search content no longer appearing in search results or being given a lower priority in the overall ranking of search results again.[1064]

6.318 These measures are part of our Illegal Content Codes for U2U and search services. The full text of the measures can be found in our Codes:

- The measure on appropriate action in determining appeals for large and/or multi-risk services is ICU D8 for U2U services and ICS D7 for search services.
- The measure on appropriate action in determining appeals for services that are neither large nor multi-risk is ICU D9 for U2U services and ICS D8 for search services.
- The measure on appropriate action following the determination of an appeal is ICU D10 for U2U services and ICS D9 for search services.

[1063] In our Codes, this measure is ICU D8 for U2U services and ICS D7 for search services.

[1064] In our Codes, this measure is ICU D10 for U2U services and ICS D9 for search services.

6.319 These measures are part of our Illegal Content Codes on terrorism, CSEA and other duties.

## Our reasoning

### How these measures work

6.320 The measures in this section outline the action providers should take in response to complaints which are appeals.

6.321 Our general position is that all providers should determine all appeals. They are an important protection for complainants' and users' rights to freedom of expression. Providers should also take appropriate steps to reverse the effects of an incorrect decision, otherwise successfully appealing does not help the complainant or user concerned.

6.322 We note requests from stakeholders that we should play a role in service providers' complaints procedures, or offer an independent appeals process. However, this is beyond the scope of our role as set out in the Act. In any event, we do not consider that such an approach would be feasible given the number of service providers in scope of the Act and the volume of complaints that providers receive.

6.323 We also do not have powers to include alternative dispute resolutions (ADR) recommendations in the Codes as currently laid out in the Act. It would be for the Secretary of State to amend the Act by regulations to include ADR following consultation with Ofcom (and others).

6.324 We note, in relation to Google's response to our November 2023 Consultation, that search service providers do not have to accept appeals relating to all downranking of content. Appeals about downranked content are only relevant where the downranking is the result of action taken in order to comply with the provider's safety duties.

6.325 Complaints about the use of proactive technology in ways that are suspected to be in breach of a provider's terms of service or publicly available statements are not appeals, and as such we have recommended different appropriate action depending on whether the service is a U2U service or a search service, as discussed in the section on 'Appropriate action for processing relevant complaints'.

#### Measure on appropriate action for determining appeals for services that are large and/or multi-risk (ICU D8/ICS D7)

6.326 This measure applies to providers of U2U services that are large or multi-risk, and providers of search services that are large general search services or multi-risk. This measure sets out the minimum targets a provider should consider when determining relevant complaints which are appeals: the time it takes to determine an appeal and the accuracy of decision making. We have not been prescriptive about the targets that providers should set as this will depend on the volume of complaints a service receives, the proportion of those complaints that are incorrectly deemed to be illegal content, and the number of subsequent appeals that emerge from these decisions. As such, providers have flexibility to set targets that are appropriate for their services.

6.327 The provider should also prepare and apply a policy about the prioritisation of appeals. This component of the measure has been amended to clarify our intention that the below factors should be considered when setting a policy for the prioritisation of appeals, not when handling the appeal. The amended component clarifies that any policy regarding the prioritisation of appeals on relevant U2U services should have regard to:

a) the seriousness of the action taken against the user or in relation to the content,
b) whether the decision that the content was illegal was made by content identification technology, and
c) the past error rate on the service in relation to illegal content judgements.

6.328 The amended component clarifies that any policy regarding the prioritisation of appeals on relevant search services should have regard to:

a) the seriousness of the action taken against the interested person,
b) whether the decision that the search content was illegal content made by content identification technology, and
c) the past error rate on the service in relation to illegal content judgements

6.329 We include the past error rate in this because it affects the likelihood that the appeal ought to be upheld. However, we do not suggest that the policy should be continually revised as the error rate changes. We only recommend that providers should consider their error rate in designing their policy. Providers are still able to take account of other factors when deciding on their policy for prioritising appeals, in addition to having regard to those specified in the measure. For example, Snap said other criteria should be used to inform prioritisation, such as the date/chronology of when an appeal was submitted.[1065] This is still possible with the measure. Our measure gives services enough flexibility to set their own prioritisation process that does not generate operational challenges whilst setting out factors that need to be considered in the creation of this process. We do not think this approach is overly prescriptive.

### Measure on appropriate action for determining appeals for services that are neither large nor multi-risk (ICU D9/ICS D8)

6.330 For providers of U2U services that are neither large nor multi-risk, and providers of search services that are neither a large general search service nor a multi-risk service, we consider it sufficient to say that appeals should be determined promptly. This measure recognises that providers of such services may have limited capacity or resources to develop and meet targets for determining appeals. Such providers are not likely to receive many complaints (and therefore fewer appeals).

### Measure on appropriate action following determination of successful appeals (ICU D10/ICS D9)

6.331 In our May 2024 Consultation, we amended this measure so that it aligned with the equivalent measure we proposed for our Children's Safety Codes. These amendments also reflected stakeholder concerns about our recommendation that successful appeals should result in content being restored to its original position. They argued it was not always appropriate or technically feasible to restore content to the exact same position, particularly in the case of ephemeral content or search results.[1066]

6.332 Since our May 2024 Consultation, we have made further amendments to the measure to clarify our intent. If a provider reverses a decision that content – or search content – was illegal content, it should:

a) Reverse the action taken.

[1065] Snap response to November 2023 Consultation, p.14.

[1066] Google response to November 2023 Consultation, p.51; Ukie response to November 2023 Consultation, p.23; Snap response to November 2023 Consultation, p.15.

i) On U2U services, this will mean reversing the action taken against the user or the content, or both. This might mean reinstating a user's account or a piece of content (where possible). On search services, this will mean reversing the action taken in relation to search content. This might mean, where possible, reinstating content that was removed from search results or removing any relevant penalty or tag that was applied to give the content a lower priority.

ii) We have amended the wording of the measure to clarify that services should only restore the content to its original position where possible. We maintain that, if a provider reverses a moderation decision following an appeal, it should also reverse the action taken as a result of that decision to the extent possible. This measure does not recommend service providers to undo any other actions they have carried out in relation to the content for reasons other than the decision that it is illegal content.

b) Adjust any relevant content or search moderation guidance if appropriate to ensure its accuracy where there is significant evidence of content being taken down, or search content not appearing in search results or being given a lower priority in the ranking of search results. In response to stakeholder feedback, we are clarifying that updates to this guidance should be made only where there is a pattern or significant evidence of errors, and only if appropriate to ensure the accuracy of the guidance.

c) Take appropriate steps to secure that the use of automated content moderation technology does not cause the same content to be taken down again, or the same search content to no longer appear in search results or be given a lower priority in the overall ranking of search results, where possible and appropriate. This works to ensure that automated content moderation technology works accurately, and that the provider ultimately makes accurate illegal content judgements. As above, we are clarifying that providers will not always need to adjust their automated content moderation technology: to do so could result in the technology not working as intended to protect users. But providers should do so where possible and appropriate to ensure that the same content is not taken down again.

## Benefits and effectiveness

6.333 We consider appeals to be an important means of protecting complainants and users against excessive takedowns of content, and in preserving their rights to freedom of expression.[1067] Determining appeals is beneficial to users as, in most cases, it enables content that has incorrectly been taken down to be restored. An appeals process is also beneficial to services as it can help them to identify and remove illegal content in the future more accurately.

### Measure on appropriate action for determining appeals for services that are large and/or multi-risk (ICU D8/ICS D7)

6.334 Appeals that are dealt with systematically and accurately can help protect rights and help build a more effective online safety. This is why we recommend providers of large or multi-risk services, who are likely to have more appeals, set targets.

6.335 We are not prescriptive about the targets that providers of large or multi-risk services should set, but recommend that these targets relate to the time taken to determine the

[1067] In their response to our November 2023 Consultation, Name withheld 5 made a similar point that effective reporting and complaints systems (including for appeals) were important to balance safety and user rights. Name withheld 5 response to November 2023 Consultation, p.15.

appeal and the accuracy of decision-making. Setting performance targets can help services be clear about the outcomes they are trying to achieve and, subsequently, measure whether they are achieving them. By monitoring their performance against these targets, we expect providers to be able to better plan, configure and refine their processes to meet their goals.

6.336 In order to ensure the measure's effectiveness, the provider will have to strike a balance between timeliness and accuracy of decision making. Providers should set their performance targets in a way that pursues both speed and accuracy of moderation and does not solely pursue one of these factors to the detriment of the other.

### Measure on appropriate action for determining appeals for services that are neither large nor multi-risk (ICU D9/ICS D8)

6.337 Our measures recognise that services differ in size and level of risk. For services that are neither large nor multi-risk, our measure set outs that providers should determine appeals promptly. We consider this to be a sufficient approach for such services as they are less likely to receive high volumes of complaints, let alone appeals.

### Measure on appropriate action following determination of successful appeals (ICU D10/ICS D9)

6.338 This measure is important to protect complainants' and users' rights to freedom of expression by ensuring that incorrect decisions that content is illegal are corrected. This helps build more effective online safety.

6.339 The measure may also indirectly encourage providers to perform a more thorough review in their initial judgments in order to avoid having to process appeals. Overall, it should increase confidence in a service provider's complaints procedures.

## Costs and risks

6.340 We consider the costs impact outlined in the preceding section ('Appropriate action for processing relevant complaints') to apply to the measures considered in this section.

6.341 In addition to this, in order for providers of large and multi-risk services to implement the measures on determining appeals, they would need to develop prioritisation frameworks and set and monitor performance targets for appeals. This would require similar activities and costs to those described in the context of Content moderation measures ICU C3 and ICU C4 and Search moderation measures ICS C2 and ICS C3. We consider that there are likely to be some overlaps in the processes needed for the content and search moderation measures and the measures on determining appeals. Any overlaps in the processes may imply lower overall costs compared to if the measures were considered independently.

6.342 The duty to consider appeals is an important way in which the Act safeguards users' rights to freedom of expression. As a result, our discretion in determining what is "appropriate" for appeals is not wide.

## Rights impact

### Freedom of expression and freedom of association

6.343 We do not consider our measures would have any negative impact on the rights of users, interested persons or services to freedom of expression or (where relevant) freedom of association. Determining appeals and taking action where possible and appropriate to reverse the impacts of any incorrect decision is an important safeguard for these rights. Taking appropriate steps where possible to ensure similar errors are not made in future is also important to help protect the rights to freedom of expression of users and interested persons.

### Privacy

6.344 We consider the rights impact on privacy as discussed in the preceding section (‘Appropriate action for relevant complaints’) to apply to the measures considered in this section.

## Who these measures apply to

6.345 While reporting and complaints duties apply to all service providers within the scope of the Act, we are making different recommendations for our measures on appropriate action for determining appeals depending on the size and risk level of the service.

6.346 We consider it likely that providers of large or multi-risk services will have to review a larger number of complaints and determine a larger number of appeals. Although this argument is clearer in the case of multi-risk services, we maintain that it is beneficial to apply measures ICU D8/ICS D7 to large services regardless of the level of risk because they have the reach and potential to affect many users.[1068] We outline our reasons for this in ‘Our approach to developing Codes measures.’

6.347 Providers of services that are not large nor multi-risk will not need to do as much to take appropriate action under this measure. For these providers we consider that it is proportionate for them to determine appeals ‘promptly’ given these services are likely to have low volumes of complaints and appeals. As such, measures ICU D9/ICS D8 will apply to services which are neither large nor multi-risk.

6.348 Our measure on appropriate action following successful appeals (ICU D10/ICS D9) will apply to all U2U and search services.

# Conclusion

6.349 Appeals are an important safeguard for the rights of complainants and others, and appropriate action in relation to them will deliver important benefits, including the restoration of content. Our measures relating to appeals will result in some costs to service providers but we consider these to be both required by the Act (in other words, we have little discretion) and marginally mitigated by overlapping measures we are recommending for search and content moderation. We consider these measures to be proportionate on the basis that service providers cannot comply with their duties under the Act without a robust appeals function. The measures we have set out in this section are necessary components of that appeals function.

6.350 Having assessed stakeholder responses, we have made changes to two of the measures considered in this section to clarify our intent. For our measure relating to determination of appeals for large and/or multi-risk services, we are clarifying what a policy for prioritisation of appeals should, at a minimum, have regard to. For our measure on action to be taken following determination of appeals of relevant complaints, we have clarified that we do not expect providers to make constant updates to their content or search moderation guidance and technology. Instead, they should do so where is significant evidence of content being taken down in error in order to avoid the same content being taken down again. We have

[1068] This means that this measure will only apply to large vertical search services if they are multi-risk services.

also clarified that we only expect a provider to restore content to its previous position where possible and appropriate.

6.351 We are including these measures in our Illegal Content Codes of Practice for U2U and search services. They are a part of our Illegal Content Codes on terrorism, CSEA and other duties. These measures are referred to as:

- ICU D8 for U2U services and ICS D7 for search services (Appropriate action for determining appeals for large or multi-risk services).
- ICU D9 for U2U services and ICS D8 for search services (Appropriate action for determining appeals for small or low-risk services).
- ICU D10 for U2U services and ICS D9 for search services (Appropriate action following the determination of an appeal).

## Measure on dedicated reporting channels for trusted flaggers for fraud

6.352 In our November 2023 Consultation, we proposed that providers of large U2U and large general search services with a medium or high risk of fraud should establish and maintain a dedicated reporting channel for specified public bodies (trusted flaggers) to use.

6.353 Section 10(3)(a) of the Act creates a duty to operate a U2U service using proportionate systems and processes that minimise the length of time for which any priority illegal content is present, with section 23(3) of the Act creating a duty for search services to minimise the risk of individuals encountering search content that is priority illegal content.

6.354 A dedicated reporting channel refers to the infrastructure and processes that enable the reporting of content and sharing of information from external organisations with service providers. Such a channel is distinct from standard user reporting. This is because a dedicated reporting channel allows for the sharing of a broader range of information than a standard user reporting channel, given the valuable intelligence and insights that expert organisations can share.

6.355 We described a trusted flagger as an entity which can offer particular expertise in notifying illegal content on a provider's service.

### Summary of stakeholder feedback[1069]

6.356 Many stakeholders from various sectors welcomed the intention behind the measure.[1070] Stakeholders provided feedback on how the measure would be implemented and how it could be strengthened, which we have grouped into the following themes:

[1069] Note: this list is not exhaustive and further responses can be found in Annex 1.

[1070] Airbnb response to November 2023 Illegal Harms Consultation, p.17; Alliance to Counter Crime Online (ACCO) response to November 2023 Illegal Harms Consultation, p.2; Association of British Insurers response to November 2023 Illegal Harms Consultation, p.2; Name withheld – a civil society organisation response to November 2023 Illegal Harms Consultation, pp.11-12; Cifas response to November 2023 Illegal Harms Consultation, p.14; [✂]; Federation of Small Businesses response to November 2023 Consultation, p.4; Innovate Finance response to November 2023 Illegal Harms Consultation, pp.13-14; International Justice Mission's Center to End Online Sexual Exploitation of Children response to November 2023 Illegal Harms

- Expanding the list of trusted flaggers for fraud.
- Selecting trusted flaggers.
- Establishing a dedicated reporting channel for fraud.
- Using standard user reporting routes.
- Functioning of a dedicated reporting channel for trusted flaggers.
- Accountability of trusted flaggers.
- Two-year review period.
- Costs.
- Rights.
- Feedback on who measure applies to.
- We detail comments on these themes in the paragraphs below.

6.357 We detail comments on these themes in the paragraphs below.

## Expanding the list of trusted flaggers for fraud

6.358 Several stakeholders called for us to broaden the list of trusted flaggers for fraud from the list of seven public bodies we proposed. They argued that greater representation of organisations with expertise in fraud would allow services to leverage this expertise, leading to benefits.[1071] The types of organisations recommended as potential trusted flaggers included additional public sector representatives (police forces, government bodies, and a regulator), consumer protection bodies, victim and survivor support groups, providers of financial services, and trade bodies representing financial services. We respond to this feedback in paragraphs 6.376-6.379 in the 'How this measure works' section.

## Selecting trusted flaggers

6.359 Several stakeholders commented on our approach to selecting trusted flaggers. One service provider sought further clarity on the criteria for Ofcom selecting trusted flaggers.[1072] Another service provider suggested that it should be at the discretion of service providers

Consultation, p.11; Match Group response to November 2023 Consultation, p.13; Monzo response to November 2023 Illegal Harms Consultation, p.16; National Trading Standards eCrime Team response to November 2023 Illegal Harms Consultation, p.11; National Trading Standards Scams Team response to November 2023 Consultation, p.2; OnlyFans response to November 2023 Consultation, p.8; TSB Bank response to November 2023 Consultation, p.10; UK Finance response to November 2023 Consultation, p.1, 7, 16; Which? response to November 2023 Illegal Harms Consultation, p.14.

[1071] Advisory Committee for Scotland response to November 2023 Illegal Harms Consultation, p.4; Association of British Insurers response to November 2023 Consultation, p.2; Name withheld – a civil society organisation response to November 2023 Illegal Harms Consultation, p.11; Cifas response to November 2023 Consultation, p.14; Innovate Finance response to November 2023 Consultation, pp.13-14; Lloyds Banking Group response to November 2023 Consultation, p.10; Logically response to November 2023 Illegal Harms Consultation, p.17; Monzo response to November 2023 Consultation, pp.16-19; [✂]; National Trading Standards eCrime Team response to November 2023 Consultation, p.11; Scottish Government response to November 2023 Consultation, p.8; The Cyber Helpline response to November 2023 Consultation, p.15; UK Finance response to November 2023 Consultation, pp.1, 7, 16; Which? response to November 2023 Consultation, pp.2, 14; TSB Bank response to November 2023 Consultation, p.10.

[1072] Vinted response to November 2023 Consultation, p.8.

to determine who trusted flaggers are.[1073] [✂].[1074] We discuss these concerns in paragraph 6.377 in the 'How this measure works' section.

## Establishing a dedicated reporting channel for fraud

6.360 Some industry stakeholders expressed concerns about the need to establish a new and separate reporting channel for trusted flaggers that was just for fraud.[1075] We address these concerns in paragraph 6.381 in the 'How this measure works' section.

## Using standard user reporting routes

6.361 Pinterest supported the EU's Digital Services Act ('DSA') approach whereby a trusted flagger would use a standard reporting route but providers would have to give reports from these trusted flaggers priority.[1076] We address this feedback in paragraph 6.382 in the 'How this measure works' section.

## Functioning of a dedicated reporting channel for trusted flaggers

6.362 Several stakeholders requested clarity regarding how a dedicated reporting channel for trusted flaggers would work in practice.[1077] This was particularly in relation to: the actions that service providers should take following receipt of a report from a trusted flagger; the speed at which service providers should take action; and the legal framework for sharing information.[1078] Google noted that trusted flaggers should include details of why the content is illegal in order to distinguish the process from a user flag.[1079] [✂].[1080] [✂].[1081] We discuss this feedback in paragraph 6.384-6.385 in the 'How this measure works' section.

## Accountability of trusted flaggers

6.363 Stakeholders requested further clarity on how trusted flaggers would be held accountable.[1082] Some stakeholders flagged the potential of misuse of dedicated reporting channels by such entities and suggested Ofcom set out how these measures would safeguard against potential abuse.[1083] One stakeholder suggested that Ofcom should 'regularly review the success of this approach' and its utility in reporting certain types of

[1073]Snap response to November 2023 Consultation, p.16.

[1074] [✂].

[1075] Meta response to November 2023 Consultation, p.29; Snap response to November 2023 Consultation, p.16.

[1076] Pinterest response to November 2023 Consultation, p.9.

[1077] [✂]; Financial Conduct Authority (FCA) response to November 2023 Illegal Harms Consultation, p.8]; Google response to November 2023 Consultation, pp.52-53; Lloyds Banking Group response to November 2023 Consultation, p.10; Meta response to November 2023 Consultation, p.29; National Trading Standards Scams Team response to November 2023 Consultation, p.2; OnlyFans response to November 2023 Consultation, p.7; Snap response to November 2023 Consultation, pp.15-16; UK Finance response to November 2023 Consultation, pp.6-7; Vinted response to November 2023 Consultation, p.8.

[1078] Airbnb response to November 2023 Consultation, p.17; [✂]; FCA response to November 2023 Consultation, p.8; Google response to November 2023 Consultation, pp.52-53; Lloyds Banking Group response to November 2023 Consultation, p.4; National Trading Standards Scams Team response to November 2023 Consultation, p.2; [✂].

[1079] Google response to November 2023 Consultation, pp.52-53.

[1080] [✂].

[1081] [✂].

[1082] Vinted response to November 2023 Consultation, p.8.

[1083] Are C. response to November 2023 Consultation, pp.7, 12; Big Brother Watch response to November 2023 Consultation, pp.4, 9; BILETA response to November 2023 Consultation, p.24.

harm.[1084] Although one service provider agreed that information from trusted flaggers is likely to be of high quality, it also suggested that content being reported by a trusted flagger does not always indicate that a harm is widespread, and therefore should not always be prioritised.[1085] A small number of stakeholders suggested that this measure creates a privileged channel for a small set of public sector organisations and flagged risks associated with this.[1086] We discuss these concerns in paragraph 6.386 in the 'How this measure works' section.

## Two-year review period

6.364 Some stakeholders raised concerns about the regularity of the review period.[1087] One stakeholder raised concerns that setting specific timeframes within which a provider should engage with trusted flaggers carried the risk of issues remaining unaddressed for too long before a review is required; instead, they suggested that concerns with the process or effectiveness of the dedicated reporting channel should be dealt with iteratively and in a timely manner when raised.[1088] We address this feedback in paragraph 6.388 in the 'How this measure works' section.

## Costs

6.365 A number of stakeholders noted that this measure could impose additional costs on providers.[1089] One stakeholder expressed concerns that the measure would require additional resources and lead to increased costs which would be disproportionate to the prevalence of the harm on the service and their existing trusted flagger mechanisms for reporting fraud.[1090] In contrast, TSB said that the proportionality concerns set out in the November 2023 Consultation are overstated.[1091] We consider this feedback in the 'Costs and risks' section.

## Rights

6.366 Some stakeholders raised concerns about the implications of the measure on users' rights, especially freedom of expression and the protection of personal data.[1092] Stakeholder responses flagged the risk of misuse of the flagging process and the need to scrutinise the use of reporting channels to ensure protection of human rights, and to prevent the removal of legitimate content or accounts. [1093] We discuss these concerns in the 'Rights impact' section.

[1084] Scottish Government response to November 2023 Consultation, p.8.

[1085] Snap response to November 2023 Consultation, p.11.

[1086] Are C. response to November 2023 Consultation, p.12; Big Brother Watch response to November 2023 Consultation, pp.4, 9; Global Partners Digital response to November 2023 Consultation, p.13.

[1087] [✂]; Monzo response to November 2023 Consultation, p.18; National Trading Standards Scams Team response to November 2023 Consultation, p.2.

[1088] FCA response to November 2023 Consultation, p.8.

[1089] Global Partners Digital response to November 2023 Consultation, p.13; Pinterest response to November 2023 Consultation, p.9; Snap response to November 2023 Consultation, p.16.

[1090] Snap response to November 2023 Consultation, p.16.

[1091] TSB response to November 2023 Consultation, p.11.

[1092] Are C. response to November 2023 Consultation, p.12; Big Brother Watch response to November 2023 Consultation, pp.4-5 and 9-10; Global Partners Digital response to November 2023 Consultation, p.13.

[1093] Are C. response to November 2023 Consultation, p.12; Big Brother Watch response to November 2023 Consultation, pp.4-5 and 9-10.

### Feedback on who this measure applies to

6.367 Some stakeholders supported the expansion of this measure to other service providers. Several civil society and consumer protection organisations recommended that the measure should be applied to all services at risk of fraud, including Small and Medium Sized Enterprises.[1094] Which? argued that the threshold for 'large' services had been set too high for fraud measures, noting that the largest mobile dating app in the UK has 2.5 million monthly users, compared with our definition of seven million users. It argued that the threshold for 'large' services should be reduced to 700,000 monthly users.[1095] We discuss these concerns in the 'Who this measure applies to' section.

## Our decision

6.368 We have decided to confirm this measure broadly as proposed in our November 2023 Consultation, with some changes:

- We have added two additional entities to the list of trusted flaggers (Police Scotland and the Police Service of Northern Ireland).
- We have clarified that providers can use their existing reporting channels for trusted flaggers (where these channels are separate from the standard user reporting process).
- We have clarified that a provider can use a dedicated reporting channel for other types of harm (and associated trusted flaggers). However, it must, at a minimum, set up a dedicated reporting channel on request, for the specified list of trusted flaggers to use to report fraud.

6.369 The full text of the measure can be found in our Illegal Content Codes of Practice, for U2U and search services and is referred to as ICU D14 for U2U services and ICS D13 for search services. This measure is part of our Illegal Content Code of Practice for other duties.

## Our reasoning

### How this measure works

6.370 This measure recommends that providers of large U2U and large general search services at medium or high risk of fraud should make a dedicated reporting channel available to any of the trusted flaggers we recommend, should they ask for access to one. The channel must be used exclusively by trusted flaggers (which could include the recommended trusted flaggers and other entities the provider has reasonably determined has expertise in a particular illegal harm). It must be distinct from standard user reporting. A service provider may also use a dedicated reporting channel for other harms should it wish to, but at minimum it should ensure one is available for reporting fraud if it is in scope of this measure.

6.371 A dedicated reporting channel is a specialised pathway for reporting harmful content – in this case, illegal content. Users of such channels (i.e. trusted flaggers) have particular expertise and competence in detecting and identifying the type of content concerned. Trusted flaggers usually represent collective interests, often through a public mandate, and operate independently from any online service.

[1094] Alliance to Counter Crime Online response to November 2023 Consultation, p.2; SPRITE+ (JN) response to November 2023 Illegal Harms Consultation, p.7; Which? response to November 2023 Consultation, p.1.
[1095] Which? response to November 2023 Consultation, pp.2-3.

6.372 The intention behind this measure is to facilitate engagement between service providers and expert organisations. Streamlined reporting channels provide a direct route for trusted flaggers to report information and intelligence to service providers. Providing a direct route for expert organisations with specific expertise in tackling fraud to contact service providers can play a valuable role in improving the detection of illegal, fraudulent content and in enabling service providers to anticipate and identify risks.

6.373 In our November 2023 Consultation, we set out the specific steps a service provider should take in relation to establishing and maintaining a dedicated reporting channel for trusted flaggers to report fraud. We are clarifying that this does not prevent the provider from making the arrangements available to other trusted flaggers who have appropriate expertise in a particular harm. The steps which we recommend should be taken, in relation to (at minimum) the specific trusted flaggers we recommend, are:

- The service provider should publish a clear and accessible policy on its processes relating to the establishment of reporting arrangements for trusted flaggers, covering any relevant procedural matters.
- If a request is made by any of the recommended trusted flaggers in accordance with the service provider's policy, the service provider should establish and maintain a reporting channel for it to use.
- The provider should engage with a recommended trusted flagger that has requested access to a dedicated reporting channel at the start of the relationship to understand its needs.
- At least every two years, the service provider should seek feedback from the recommended trusted flaggers, on whether any reasonable adjustments or improvements might be made to the operation of the reporting channel.

6.374 A complaint from a trusted flagger will amount to reason to suspect the content is illegal for the purposes of ICU C1 or ICS C1. As set out in ICU/ICS A5 (tracking evidence of new and increased harm), reports from trusted flaggers about other matters will be relevant to a provider's ongoing risk monitoring and management process.

6.375 These steps describe how we expect providers to engage with recommended trusted flaggers. A service provider may choose to apply those same steps in its engagement with other organisations.

### Expanding list of trusted flaggers for fraud/selecting trusted flaggers

6.376 Considering feedback from the November 2023 Consultation, we have added two new entities to the list of recommended trusted flaggers. Our list of trusted flaggers now contains nine public bodies instead of seven. The two additional trusted flaggers we have included will ensure full representation across the UK, in line with our original policy intention of creating a safer life online for all UK users. The City of London Police, as the lead force for fraud in England and Wales, provides coverage for both those nations.

6.377 We have recommended this list of trusted flaggers given their expertise in detecting and tackling fraud. These organisations represent collective interests, often through a public mandate, and operate independently from service providers. They bring enhanced credibility in the reporting process and may provide valuable information regarding the prevalence of fraud on services.

6.378 Our list of nine trusted flaggers, and our justification for including them, includes:

- (new) **Police Scotland** and the **Police Service of Northern Ireland** to ensure full representation across the UK.
- The **City of London Police** is the national lead police force for fraud and cyber security for England and Wales.
- The **National Economic Crime Centre**[1096] and **National Crime Agency** coordinate a multi-agency system response to economic crime and play essential roles in understanding the changing nature of fraud online. The National Economic Crime Centre coordinates the UK's response to economic crime, harnessing intelligence and capabilities from across the public and private sectors.
- The **National Cyber Security Centre** operates an online reporting portal for organisations and individuals to report scam website links and URLs, many of which are also likely to be shared or promoted by fraudsters via user-generated content posted on online services.
- The **Dedicated Card and Payment Crime Unit** (a joint partnership between law enforcement and the banking sector) has partnered with several social media platforms to identify accounts that feature posts linked to payment crime.
- The **Financial Conduct Authority** has useful insights to share with online service providers in relation to investment scams and financial promotions scams. The Financial Conduct Authority is also in a position to provide important information and expertise on certain financial-services-related priority offences.
- Certain Government departments have a particular interest and expertise in respect of fraud. **HM Revenue and Customs** and the **Department for Work and Pensions** each have a large customer base and are closely sighted on emerging trends relating to fraud that targets people by reference to matters relating to their tax or benefits (as the case may be).

6.379 Further expansion of the list of trusted flaggers would depend on an assessment of the proportionality of doing so, including the costs and resource implications and the benefits to users. We may consider expanding the list to include further organisations in the future.

#### Establishing a dedicated reporting channel for fraud and using standard reporting routes

6.380 For the purposes of reporting fraud, the nine trusted flaggers we are recommending should be considered a minimum. As explained in paragraph 6.373, a service provider does not need to establish a relationship with all nine trusted flaggers, but should do so for any that request access to a dedicated reporting channel. We recognise that a service provider may have other, pre-existing trusted flagger relationships – for the purpose of reporting fraud and other harms – with organisations not listed in our Codes., A service provider may also want to utilise other trusted flaggers beyond the ones we are recommending as part of this measure. It is at the service provider's discretion to continue these pre-existing relationships or create new reporting relationships with trusted flaggers beyond the ones we have listed.

6.381 Two service providers sought clarity over whether existing dedicated reporting channels for trusted flaggers could be used, or if providers would be expected to create a new and

[1096] The NCA's National Economic Crime Centre (NECC) is a multi-agency centre that was established to deliver a step-change in the response to tackling serious and organised economic crime.

separate channel exclusively for fraud.[1097] We have amended our measure to be clearer that an existing dedicated reporting channel for trusted flaggers may be used where it meets the steps set out in the Codes. We are clarifying that trusted flaggers must have a distinct route separate from that of standard complainants to communicate with a service provider.

6.382 We are also clarifying that a dedicated reporting channel for trusted flaggers does not need to be used solely for the purpose of reporting fraud and can be used by a provider to tackle other harms raised by other trusted flaggers. In this measure, we are setting out that, at a minimum, a large U2U service provider at high or medium risk of fraud and a large general search service provider at high or medium risk of fraud must have a dedicated reporting channel that recommended trusted flaggers can use to report fraud; this does not prevent a service from using the dedicated reporting channel for reports and complaints about other types of harm from other trusted flaggers.

### Function of a dedicated reporting channel for trusted flaggers

6.383 The design of a reporting channel and is a matter for the service provider to determine. It is for the trusted flagger concerned to determine what information it can lawfully share. A service provider can engage with trusted flaggers to ascertain how best to design a dedicated reporting channel for reporting fraud. These considerations may evolve over time as providers engage with trusted flaggers to understand their reporting needs and as the harm evolves.

6.384 In terms of how quickly providers should review and take action on reports, this should be determined in accordance with the provider's content moderation policies, as set out in our Codes as ICU C3/ICS C2. We expect that the service provider will handle the complaints from trusted flaggers received through the dedicated reporting channel through its content or search moderation function.

6.385 Under those measures, it is the responsibility of the service provider to prepare and apply a policy regarding the prioritisation of content for review. In setting the policy, the provider should have regard to several factors, including whether the content has been reported by a trusted flagger. This is set out in chapter 3 of this Volume: 'Search Moderation'.

### Accountability of trusted flaggers

6.386 We have reflected on feedback regarding the performance and accountability of trusted flaggers. We have also reflected on the potential for misuse and the risks associated with creating a privileged channel for UK Government bodies and law enforcement. We consider the risk to be minimal for the set of trusted flaggers we are proposing given it consists of public entities with expertise and competence in relation to tackling fraud. They are, by definition, subject to duties to act fairly and proportionately and not to act in a way which is incompatible with human rights. It is our expectation that these trusted flaggers will utilise dedicated reporting channels in an appropriate manner.

6.387 A service provider should raise any concerns about the use of a dedicated reporting channel by a trusted flagger with that trusted flagger in the first instance. The provider may also choose to establish an escalation and dispute resolution process as a part of its dedicated reporting channel policy.

[1097] Meta response to November 2023 Consultation, p.29; Snap response to November 2023 Consultation, p.16.

### Two-year review period

6.388 We have considered the feedback regarding the two-year review period. We consider that a defined review period is necessary to ensure that trusted flaggers re-engage with service providers. We consider that a minimum two-year review period provides service providers with a sufficient amount of time to assess the efficacy of the arrangements. A more frequent review period may not allow service providers and trusted flaggers sufficient time to assess the impact and effectiveness of the dedicated reporting channels.

## Benefits and effectiveness

6.389 The need for this measure is evidenced by the scale of harm occurring from fraud (both online and offline), which is the most frequently experienced crime in the UK. It currently accounts for almost two fifths (38%) of all crime in England and Wales and has more than doubled in Scotland over the past nine years.[1098] Four fifths (80%) of reported fraud is cyber-enabled and the use of social media and encrypted messaging services as an enabler is increasing throughout all aspects of fraud.[1099] Nine out of ten adult internet users (87%) have encountered content online which they believed to be a scam or fraud, demonstrating the scale of this threat.[1100]

6.390 While it is challenging to estimate the economic and social cost of fraud, the UK Government Fraud Strategy estimated it to be £6.8bn in England and Wales in 2019 to 2020.[1101] The average cost per fraud incident was estimated to be £1,427.[1102] Our Register chapter titled 'Fraud and Financial Services' sets out in more detail the scale of the harm caused by online fraud.

6.391 We consider that this measure will deliver benefits for providers and users. The measure provides the building blocks for providers to obtain useful intelligence regarding fraud on their services. This information should empower the provider to detect and tackle the issue of fraud, and limit user exposure. This measure will help combat fraud and deliver benefits to users by enabling providers to leverage intelligence from expert organisations. For example, the law enforcement organisations listed as recommended trusted flaggers in this measure can send information and alerts to relevant providers regarding fraud. Intelligence regarding the use of services to enable fraud will contribute to a provider's efforts to tackle the prevalence of this harm and protect users online.

6.392 The creation of a dedicated reporting channel will make providers materially more likely to identify and remove fraudulent content on the services they operate. As explained in our November 2023 Consultation, we therefore expect this measure to make a meaningful contribution to reducing this harm.

6.393 A distinct channel for trusted flaggers ensures that more advanced intelligence and evidence can be shared even if it is not directly tied to a specific piece of content. This intelligence and evidence may relate to the broader dynamics of how fraud manifests on the service. Such information can help the service provider understand risks and emerging harms.

[1098] Home Office, 2024. Fraud Factsheet, [accessed 10 November 2024].

[1099] National Fraud Intelligence Bureau (NFIB), 2020/2021. Fraud Crime Trends. [accessed 16 September 2024].

[1100] Ofcom, 2023. Online Scams and Fraud, [accessed 16 September 2024].

[1101] Home Office, 2023. Fraud Strategy: Stopping Scams and Protecting the Public, [accessed 16 August 2024].

[1102] Department of Digital, Culture, Media and Sport, 2022. Online Safety Bill – Impact Assessment, [accessed 4 September 2024].

6.394 The measure also reflects current efforts by providers to streamline reporting. It aligns with the voluntary commitment already made by a number of providers of large services in the Online Fraud Charter to establish dedicated liaisons (points of contact) for reporting fraud. These liaisons will respond to law enforcement requests. [1103]

## Costs and risks

6.395 This measure involves the service provider developing a policy on its processes relating to trusted flagger reporting arrangements. If a request is made by one of the trusted flaggers listed, the provider would need to engage with it and provide it with access to a dedicated reporting channel. If the provider does not have a dedicated reporting channel for trusted flaggers in place, it will need to design and implement a new one. Any initial start-up costs will largely be one-off costs and will not vary greatly with the size of the service provider. However, there will be ongoing costs involved in maintaining the channel.

6.396 If the provider already has a suitable dedicated reporting channel for trusted flaggers, it may offer access to this existing channel as a starting point, so long as the channel meets the criteria set out in the Codes. This will tend to lower costs for those providers compared to if the measure recommended that they establish new channels.

6.397 If a service provider develops a clear reporting process and uses that in a consistent way when dealing with any trusted flagger, it should result in cost efficiency when developing and maintaining the reporting function and when engaging with trusted flaggers.

6.398 We originally proposed seven trusted flaggers, but are now recommending nine. We acknowledge that this may increase costs given the need to engage with two additional organisations. However, we do not consider these costs to be disproportionate given that the additional law enforcement entities are similar in nature to the ones already specified in the list, but have specific expertise in relation to particular parts of the UK. In addition, the provider would only need to engage with a recommended trusted flagger if it requests access to a dedicated reporting channel.

6.399 The organisations we have listed are all public bodies. We expect that each one has an incentive to seek a clear method that it can use regularly and consistently to share its particular expertise and competence for detecting and identifying illegal content with service providers covered by this measure.

6.400 We expect the reports provided by trusted flaggers through a dedicated reporting channel to be clearly targeted at fraudulent content. Where a service provider experiences high ongoing costs resulting from a large number of reports from relevant trusted flaggers, there are also likely to be correspondingly high benefits for users in the form of a reduction in their exposure to fraud-related content. We are therefore less concerned about increases in such ongoing costs, as they are likely to be proportionate given the benefits. Conversely, providers that adopt these measures but have low volumes of fraud-related reports are unlikely to have high ongoing costs.

6.401 Providers of services that are also subject to the relevant part of the DSA will already be required to establish trusted flagger schemes. Where there is an overlap between that requirement and our measures, additional costs will be less for providers subject to both.

[1103] Home Office, 2023. Online Fraud Charter, [accessed 23 September 2024].

6.402 As fraudulent activity evolves, and as trusted flaggers obtain different insights into fraud related to online services, the nature of the reporting channel may evolve. This may result in growing costs over time.

## Rights impact

### Freedom of expression and freedom of association

6.403 We received feedback highlighting the measure's potential impact on users' freedom of expression. We consider this impact to be minimal as the measure does not automatically recommend a service provider to take down content. However, we recognise that a service provider can determine how it implements the trusted flagger process.

6.404 There is also the possibility of competing interests between the service provider and trusted flagger. While the service provider has a duty to have particular regard to protecting users' right to freedom of expression, there is no specific requirement for trusted flaggers to do this under the Act. However, all the entities we are recommending as trusted flaggers are public bodies which are subject to their own duties in relation to fairness, proportionality and not acting incompatibly with human rights. We recognise that this does not necessarily remove the risk to human rights entirely, but we consider that it does act as a safeguard.

6.405 Fraud is one of the more difficult kinds of offence for a provider to identify online. The trusted flaggers we are recommending have expertise and access to intelligence about fraud which we consider is likely to make a material difference to service providers' ability to protect users from harm, and which is unlikely to be possible to make available to service providers in any other way. The measure therefore represents the least intrusive means of mitigating the harm. Therefore, weighing up the importance of protecting users from crime against the possible impacts on human rights that this measure will cause, we consider that the interference is proportionate.

### Privacy

6.406 We consider that this measure will have limited negative impact on users' (and others') rights to privacy. Service providers have a duty to have particular regard to the importance of protecting users' rights to privacy, including data protection, when implementing complaints processes. If done correctly, the risk of interference with these rights is minimised.

### Data protection

6.407 In implementing a trusted flagger process or utilising an existing process that aligns with the recommendations set out in the Codes, the service provider should comply with their obligations under data protection laws and familiarise themselves with any relevant guidance from the ICO.

6.408 Regarding data protection, information shared through the dedicated reporting channel (in the form of complaints regarding specific content) should be handled in accordance with the providers' processes. Service providers should ensure that reports from trusted flaggers are handled in a way that complies with data protection laws.

## Who this measure applies to

6.409 In the November 2023 Consultation, we considered it proportionate to recommend this measure for providers of large U2U services and large general search services that have identified a medium or high risk of fraud in their most recent risk assessment. Providers of such services are likely to be sufficiently resourced and able to absorb the costs of this measure. Also, the benefits may be higher as fraud is a volume crime, so more users may be helped on such services. We have not altered our view on this following stakeholder feedback, and maintain that this measure should apply to large U2U and large general search services at medium or high risk of fraud.

6.410 As noted in paragraph 6.367, Which? argued that the threshold for 'large' services had been set too high for fraud measures.[1104] Based on the information available to us at present, we are not confident it would be proportionate to apply this measure to smaller services at this stage. The benefits of applying the measure to smaller services are likely to be smaller (because of their lower reach) and there is a significant fixed element to the costs, though we do not currently have a good estimate of the scale of this.

6.411 We recognise that limiting the measure to large services creates a risk of displacement of the harm to smaller services. Even if this were to happen, the harm to users is still likely to be reduced given the smaller greater reach of larger services. As we gain a better understanding of the costs and benefits of this measure in the future, we can review the services it applies to.

## Conclusion

6.412 There are significant harms to UK users from fraud. The trusted flaggers we have identified are well placed to accurately identify fraudulent content earlier than it would otherwise be discovered. This supports providers in taking action on content more quickly, lowering the risk of users being exposed to it, and aids intelligence-sharing more broadly. We therefore anticipate this measure delivering significant benefits to users and service providers, making a meaningful contribution to reducing harm from fraud.

6.413 We have not been able to quantify the costs of the measure given the broad range of service providers in scope of it and the variety of different ways they could implement it. Nonetheless, we are targeting the measure at large U2U and large general search services that will generally be well-resourced. None of the submissions we received in response to the consultation provided clear evidence that the measure would be disproportionately burdensome for providers of such services. Given this, the important benefits a dedicated reporting channel would have, and the strategic importance of combatting fraud, we consider that the measure is proportionate.

6.414 Following stakeholder feedback, we have expanded the list of recommended trusted flaggers to include the Police Service of Northern Ireland and Police Scotland. This will ensure full representation across the UK nations and will be in line with our original policy intention of creating a safer life online for UK users. We have also addressed concerns regarding proportionality by explaining that service providers may use an existing dedicated reporting channel for trusted flaggers, provided the channel meets the criteria set out in the Codes.

6.415 We are including this measure in our Illegal Content Codes of Practice for other duties, for U2U and search services. This measure is referred to as ICU D14 for U2U services and ICS D13 for search services.

[1104] Which? response to November 2023 Consultation, pp.2-3.

# 7. Recommender systems

## What is this chapter about?

Content recommender systems are algorithmic systems used to curate personalised feeds of user-generated content and to aid the organic discovery of such content. The evidence we have suggests that these systems can play a role in increasing the risk of users encountering certain types of illegal content. Through the responsible monitoring of these systems, service providers can manage some of this risk.

This chapter discusses steps service providers can take to help them better understand the risks their content recommender systems pose. The chapter presents a measure designed to give service providers a methodical way of monitoring the risk of ongoing design adjustments to their content recommender systems by collecting safety metrics when they carry out on-platform testing.

## What decisions have we made?

We are recommending the following measure:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| ICU E1 | Providers should, when carrying out on-platform testing of content recommender systems, **collect additional safety metrics when making design adjustments**. | Providers of U2U services that:<br>• carry out on-platform testing of their content recommender systems; and<br>• are at medium or high risk of two or more specified kinds of illegal harm. |

## Why have we made these decisions?

Many service providers carry out on-platform tests when they are making adjustments to the design of their recommender systems. Typically, these focus on the impact that design adjustments have on user engagement with the service. We are recommending that service providers incorporate safety metrics into their on-platform tests. This will help providers better understand whether a design adjustment to their recommender systems might contribute to illegal content risk and, if so, how and why. This will enable them to make more informed choices about the design of their content recommender systems, and be better placed to manage risks associated with these algorithms. This should help reduce the amount of illegal content disseminated by content recommender systems.

## Introduction

7.1 To meet their illegal content safety duties in the Act, user-to-user ('U2U') service providers must use proportionate measures to mitigate and manage the risks of harm to individuals identified in their most recent risk assessment, which includes an assessment of  the risk of users encountering illegal content by means of algorithms used by the service.[1105] The safety duties require providers to take measures that relate to the "design of functionalities, algorithms and other features" and "risk management arrangements", if proportionate to do so.[1106] In this chapter, we focus on steps providers can take to improve user safety in relation to their content recommender systems and underlying algorithms.

7.2 Content recommender systems are a form of artificial intelligence ('AI') used to curate personalised content feeds on U2U services and aid the organic discovery of content from multiple users. These systems can help connect content creators with their audiences and help users encounter content that they are likely to enjoy.

7.3 Content recommender systems are comprised of many algorithms. These are sets of computing instructions that use multiple factors to determine the content shown to a user. Advanced recommender systems often use machine learning ('ML') techniques to observe and learn about a user's behavioural patterns in relation to content, enabling them to make relevant content recommendations to achieve engagement targets. Despite their benefits, content recommender systems are not without risks. We have set out evidence in our Register of Risks ('Register') that demonstrates that they can contribute to increasing the organic reach of illegal and harmful content.[1107]

7.4 References to 'recommender systems' throughout this chapter should be understood to refer only to content recommender systems (subject to the clarifications outlined in paragraph 7.8).[1108]

## Measure on the collection and monitoring of safety metrics during on-platform testing of recommender systems

7.5 In our November 2023 Illegal Harms Consultation ('November 2023 Consultation'), we proposed that, when undertaking existing on-platform tests, service providers should collect safety metrics that will allow them to evaluate whether their proposed changes to the design of a recommender system are likely to increase user exposure to illegal content. We recommended that providers record the safety metrics (and other information relevant to test results) in a log, which should be made available to relevant staff and referred to when making future design changes.

7.6 We proposed this measure because gathering information about the impact of recommender system design changes on the dissemination of illegal content will put services in a position to make materially better-informed design choices than they otherwise would. We expected this would reduce the risk of users organically encountering illegal content by means of the recommender systems and suffering harm as a result. For the purposes of this

[1105] Section 9(5)(b) and 10(2)(c) of the Act.
[1106] Section 10(4)(a) and (b) of the Act.
[1107] Please refer to the Illegal Harms Register of Risks and the draft Children's Register of Risks [accessed 8 October 2024] for a discussion of the risks associated with recommender systems, both in general and in relation to each kind of illegal harm and kinds of content harmful to children.
[1108] Examples of content recommender system channels include personalised newsfeeds, reels, 'for you' pages, and 'discover' pages.

measure, we did not intend to capture content recommender systems employed in the operation of search functionalities on a U2U service.

## Summary of stakeholder feedback[1109]

7.7 In their responses to the November 2023 Consultation, feedback from stakeholders was mixed. Non-industry stakeholders including the National Society for the Prevention of Cruelty to Children ('NSPCC'), the Childrens Commissioner, the Integrity Institute, the Institute for Strategic Dialogue, the Oxford Disinformation and Extremism Lab, and the Betting and Gaming Council were broadly supportive of this measure.[1110] Below, we outline the other key themes we have identified from our analysis of stakeholders' responses.

- **Interaction with risk assessment duties** – Google noted that the risk assessment duty relating to 'significant change' under the Act would require an assessment of user risk associated with design changes and that there is no safety justification for Ofcom to recommend other assessments in the Illegal Content Codes of Practice ('Codes') which are triggered by any lower threshold of change in recommender system design.[1111] Google also noted its understanding that all changes to recommender systems are potentially in scope of the measure. We address this issue in the 'How this measure works' section.
- **Frequency of collecting safety metrics** – LinkedIn suggested that since many adjustments to recommender systems are not ultimately deployed, there should only be periodic assessments of safety metrics (annual, biannual, or quarterly).[1112] We address this issue in the 'How this measure works' section.
- **Cost of complying** – Given the frequency of incremental changes/adjustments made to recommender systems, Google suggested that our measure would create an "enhanced compliance burden" for on-platform testing.[1113] We address this issue in the 'Cost and risks' section.
- **Flexibility** – Service providers highlighted the importance of allowing flexibility in how they approach safety-oriented evaluations of their recommender systems.[1114] These providers queried whether alternative approaches to safety testing would qualify them as compliant with the relevant safety duties. Google additionally noted the challenges of isolating UK user complaints about potential breaches relating to illegal content types covered by the Act from larger datasets of terms of service breaches.[1115] We address this issue in the 'How this measure works' section.

[1109] Note this list is not exhaustive – further responses can be found in Annex 1.

[1110] Betting and Gaming Council response to November 2023 Illegal Harms Consultation, p.11; Childrens Commissioner's response to the November 2023 Illegal Harms Consultation, p.23; Institute for Strategic Dialogue response to November 2023 Illegal Harms Consultation, p.12; Integrity Institute response to November 2023 Illegal Harms Consultation, p.17-19; National Society for the Prevention of Cruelty to Children (NSPCC) response to November 2023 Illegal Harms Consultation, pp.39-40; Oxford Disinformation and Extremism Lab response to November 2023 Illegal Consultation, p.16.

[1111] Google response to November 2023 Illegal Harms Consultation, pp.57-58.

[1112] LinkedIn response to November 2023 Illegal Harms Consultation, p.14.

[1113] Google response to November 2023 Consultation, pp.57-58.

[1114] Google response to November 2023 Consultation, pp.57-58; Meta response to November 2023 Illegal Harms Consultation, p.32.

[1115] Google response to November 2023 Consultation, p.20; Google response dated 15 July 2024 to our follow-up email dated 24 June 2024.

- **Who this measure applies to** – Snap, Meta, and Google said that our decision on who the measure applies to should not disincentivise on-platform testing and should encourage the responsible operation of recommender systems while ensuring a level playing field between service providers.[1116] The NSPCC suggested that all large multi-risk services, and not just those already conducting on-platform testing, should apply the measure.[1117] We address these points in 'Who this measure applies to' section.

- **Enforcement and accountability** – The Molly Rose Foundation (MRF), the NSPCC, and the Integrity Institute expressed concern that our proposal would not sufficiently incentivise services to mitigate high-risk design adjustments uncovered during testing.[1118] Additionally, the Integrity Institute argued that the measure creates no accountability because service providers are not required to act upon the safety metrics in a specific way, for example by adopting the safer variant of the recommender system of those tested (other than to refer to the metrics on an ongoing basis).[1119] We address this point in the 'Using the metrics as a means of identifying and managing risk to users' section.

- **Types of content recommender systems in scope** – [✂] and Booking.com submitted that it would be disproportionate to impose requirements on service providers to collect safety metrics where it has determined that their recommender system does not materially affect the likelihood or impact of the risks posed by illegal content.[1120] Separately, and in respect of Ofcom's risk assessment, Google said that there is no evidence that the content recommender system employed by Google Photos would increase risk to users on the basis that it recommends only photos from that user's gallery.[1121] We clarify these points in the 'How this measure works' section.

- **Personal data processing** – the Information Commissioners Office (ICO) highlighted that this measure could require the processing of personal data as part of the safety metrics and suggested that we update our assessment to suggest that providers adopting this measure should ensure that they comply with the purpose limitation and data minimisation principles.[1122] We address this point in the 'Data protection' section.

- **Lack of evidence of harm** – The Cyber Threats Research Centre at University of Swansea's response highlighted that there is not sufficient research that suggests the widespread proliferation of illegal content by recommender systems.[1123] We address this issue in the 'Benefits and effectiveness' section.

[1116] Google response to November 2023 Consultation, pp.57-58; Meta response to November 2023 Consultation, p.32; Snap response to November 2023 Illegal Harms Consultation, p.22.

[1117] NSPCC response to November 2023 Consultation, pp.39-40.

[1118] Integrity Institute response to November 2023 Consultation, pp.17-19; Molly Rose Foundation response to November 2023 Illegal Harms Consultation, pp.32-33; NSPCC response to November 2023 Consultation, pp.39-40.

[1119] Molly Rose Foundation response to November 2023 Consultation, pp.32-33.

[1120] [✂]; Booking.com response to November 2023 Illegal Harms Consultation, p.19; Evri response to the November 2023 Illegal Harms Consultation, p.8-9.

[1121] Google response to November 2023 Consultation, p.9.

[1122] Information Commissioner's Office (ICO) response to November 2023 Illegal Harms Consultation, p.20.

[1123] Cyber Threats Research Centre, Swansea University response to November 2023 Illegal Harms Consultation, p.12.

## Our decision

7.8 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation. In response to stakeholder feedback, we have made minor changes, clarifying the design changes and the types of content recommender systems that are relevant for this measure:

- We have replaced the term ‘design change’ with ‘design adjustment’ to delineate between significant changes that are subject to the risk assessment duty and the more incremental adjustments that are in scope of this measure.
- We have further clarified the scope of content recommender systems covered by this measure.[1124] The measure excludes product recommender systems that are used exclusively for the purpose of recommending goods and services. The measure also excludes content recommender systems that recommend only content uploaded or shared by a single user.
- For the avoidance of doubt, we have further clarified that this measure does not apply to content recommender systems that are employed exclusively in the operation of a search functionality and which suggest content to users in direct response to a search query.
- We have clarified that by ‘log’, we mean a form of record keeping that enables the continuous collection, storage and analysis of information relevant to the operation of algorithmic systems.

7.9 The full text of the measure can be found in our Illegal Content Codes of Practice for User-to-User services, in which it is referred to as ICU E1. This measure is part of our Codes on terrorism, child sexual exploitation and abuse (‘CSEA’) and other duties.

## Our reasoning

### How the measure works

7.10 In between making significant changes, service providers make smaller, more frequent design adjustments to their live recommender systems.[1125] Although such adjustments are not significant from a design perspective, they are important updates to ensure the system is performing in a way that is optimal for both the service and for end-users. Service providers may refer to them as in-production updates or off-cycle updates. While frequent design adjustments to recommender systems are essential to delivering an optimal user experience, they may, under certain circumstances, play a role in the amplification of illegal content.[1126]

[1124] We note that Government has laid before Parliament a statutory instrument under Schedule 11 to the Act (due to come into force next year) which includes a definition of ‘content recommender system’ for the purposes of a threshold condition for Category 1 services. The definition that we have employed for the purposes of this measure of content recommender system varies from the definition adopted by Government to ensure that it reflects our policy intention as described in this chapter, which has taken into consideration relevant feedback received in response to our November 2023 Consultation.

[1125] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content. [accessed 8 October 2024].

[1126] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content. [accessed 8 October 2024].

7.11 Many U2U service providers, particularly the largest, evaluate the effect of design adjustments through on-platform tests. Through this type of testing, service providers can observe the performance of the recommender system in real-time against relevant metrics. Commonly used on-platform testing methods include:

- **A/B testing** – This is a randomised controlled trial in which a test group of users (known as a treatment group) is served content from the adjusted recommender system, while a control group of users continue to be served content from the current recommender system. The results are then compared and used to decide whether to implement the design adjustment.
- **Multi-arm bandit (MAB) testing** – Unlike A/B testing, which uses static treatment and control groups, MAB testing is a continuous experiment that uses ML techniques to dynamically allocate users to the best-performing variant of a recommender system against a particular metric (such as average click-through rate per user) based on real-time data being gathered during the test. The aim of MAB testing is to quickly learn which variant of the recommender system is performing optimally, then to gradually allocate all users to it.

7.12 These tests highlight the likely effects of design adjustments on a recommender system's key performance metrics. We understand that on-platform testing typically focuses on user engagement metrics (such as the number of likes, shares, advert clicks, and time spent on a service), enabling the service provider to operate the recommender system in a commercially optimal way.

7.13 Based on our understanding of the effectiveness of these tests (outlined onwards from paragraph 7.40), to implement the measure, service providers will need to incorporate additional safety metrics into their existing on-platform tests of certain content recommender systems to evaluate whether design adjustments are likely to increase user exposure to illegal content. The following steps should be implemented to achieve this:

- Service providers should produce safety metrics to understand the potential effects of a recommender system design adjustment on the dissemination of illegal content. Our recommended safety metrics for providers are outlined in the 'Safety metrics' section.
- Service providers should keep a log, recording all test results, a description of the design adjustment, the safety metrics produced, and an explanation of the decision that was taken at the end of the test. Logs should give a reasonable indication of design adjustments that contribute to an increase or decrease in the dissemination of illegal content.
- The log should be made available to staff involved directly or indirectly in the development and testing of recommender systems (such as engineering and trust and safety teams) and should be referred to when making future adjustments.

### Distinguishing recommender system design adjustments from significant changes

7.14 As outlined in paragraph 7.7 above, referencing the requirement in the Act to carry out a further risk assessment relating to the impact of significant changes to the design or operation of a service, Google said that there is no safety reason, nor any justification under the Act for Ofcom to effectively lower the threshold of significant change by recommending our proposed measure. Google argued that the risk will have already been assessed at the point of significant design change in line with the risk assessment duties.[1127] To clarify, the risk assessment duty under section 9(4) of the Act relating to significant changes is distinct from the safety duties under section 10(2) of the Act, which require service providers to use proportionate measures relating to the design and operation of the service to effectively mitigate and manage the risk of harm to users presented by illegal content on the service.[1128]

7.15 The safety duty is clear that both the design of algorithms (such as recommender systems) and risk management arrangements are among the types of measure that providers must adopt, where proportionate (and consequently that we may recommend where proportionate).[1129] Risk management associated with the design of algorithms is an ongoing activity, not an activity triggered only by a significant change. The process of managing a recommender system is one involving frequent design adjustments which, individually, may not amount to a significant change (and so would not trigger a risk assessment), but cumulatively could leave users exposed to increased risks over time if left unmonitored and unmanaged. As outlined in the Register, there is clear evidence suggesting that content recommender systems may play a role in disseminating illegal content where it is shared on a service, and that providers make changes (including design adjustments covered by this measure) to the design of these systems in a way that may result in users being more likely to be exposed to illegal content.[1130]

7.16 We are therefore concerned with the risk of providers making these smaller, more incremental changes in a way that results in users being more likely to be exposed to illegal content, but without understanding the likely risk to users.

7.17 Our intention is therefore that this measure would apply in respect of these more frequent adjustments to the design of recommender system. In paragraphs 7.33 and 7.34 below, we explain in greater detail what we consider to be covered by our reference to 'design adjustment', and how this differs from changes to a recommender system that we would consider to be 'significant' for the purposes of the risk assessment duty.[1131] By taking steps to improve awareness of the risks associated with recommender system design, we expect this measure to improve both the ability of providers to manage those risks and the safety of design choices in relation to recommender systems. Ultimately, this may lead to a reduction in user exposure to illegal content and the risk of harm associated with that. We therefore consider that this measure presents a proportionate and effective means by which service providers may comply with the illegal content safety duties under section 10(2) of the Act.

### Content recommender systems

7.18 This measure applies to design adjustments made to "content recommender systems", by which we mean an algorithmic system that curates personalised content feeds on U2U services and aids the organic discovery of content. These systems can help connect content creators with their audiences and help users encounter content that they are likely to enjoy. In our November 2023 Consultation, we explained that this measure would not apply to

[1127] Google response to November 2023 Consultation, p.9.

[1128] Section 10(2)(c) of the Act.

[1129] Section 10(4)(a) and (b) of the Act.

[1130] For an overview of the role recommender systems play in increasing risks see the Register of Risks chapter 'Governance, systems and processes'. See also individual harms chapters in the Register of Risks for information regarding the link between recommender systems and each particular kind of illegal harm.

[1131] See Risk Assessment Guidance for further information on this duty.

content recommender systems that underpin search functionalities on a U2U service. As set out in paragraph 7.8, we have further clarified that this measure does not apply to content recommender systems that are employed exclusively in the operation of a search functionality and which suggest content to users in direct response to a search query. Increasing levels of integration between content recommender systems and search functionalities embedded within user-to-user services means that the two features can be powered by a single algorithmic system. Where this is the case, we would expect the content recommender system to be in scope of this measure notwithstanding the underlying relationship with the search functionality. This measure also does not apply to network recommender systems that recommend other users to connect with or groups to join.

7.19 In response to our November 2023 Consultation, [✂] and [✂] indicated that the risk posed by recommender systems is dependent on the purposes for which a recommender system has been implemented. Where the content being recommended is unlikely to be illegal or harmful (e.g. travel accommodation options), the risks posed by the recommender system are likely to be materially reduced.[1132] We recognise that there are different types of recommender systems, and that the level of illegal content risk may differ. Based on this feedback, our Register has been updated to reflect that references to 'content recommender system' do not include those that are designed to exclusively recommend goods and services posted by users, such as those used by online U2U marketplaces (referred to as 'product recommender systems').[1133]

7.20 In line with this change in our Register, this measure would therefore not apply to product recommender systems. We have clarified this through an amendment made to the definition of 'content recommender system' in the Codes.

7.21 As outlined in paragraph 7.7, Google separately suggested that Ofcom's Register should distinguish between different types of content recommender systems and in particular, argued that there is no evidence that those used to suggest a user's own content from their private inventories (for example, on Google Photos) present an illegal content risk.[1134] We understand that photo sharing and storage services may use private/closed content recommender systems to curate personalised 'albums' based on specific themes (e.g., revisiting a particular period) which would be limited to a user's private content. Based on this feedback, we agree that content recommender systems that recommend content from a single user would not pose a significant illegal content risk, and our Register has therefore been updated to reflect those references to 'content recommender systems' refer to recommender systems that curate content from multiple users.[1135]

7.22 In line with this change in our Register, this measure would therefore not apply to content recommender systems that only recommend a user's own content from their own private inventory. This has been clarified through an amendment made to the definition of 'content recommender system' in the Codes.

### Safety metrics

7.23 The measure recommends that service providers should collect the following safety metrics when they decide to run on-platform tests to measure the impact of design adjustments on

[1132] [✂].

[1133] Register of Risks: 'Glossary' chapter

[1134] Google response to November 2023 Consultation, p.9.

[1135] Register of Risks: 'Glossary' chapter the performance of their recommender systems. This should then enable the provider to understand whether a design adjustment would increase the risk of users encountering illegal content compared to the existing variant:

- **Total number of content items identified as illegal content (or as an illegal content proxy[1136]) during testing.**
- **Total impressions and reach per item identified as illegal content (or as an illegal content proxy)** – 'Impressions' refers to the total number of times that each piece of content identified as illegal content (or an illegal content proxy) was encountered (meaning viewed or interacted with) by users. This count includes multiple encounters by the same user. 'Reach' refers to the total number of unique users who encountered each piece of content identified as illegal content (or an illegal content proxy).

7.24 While we recommend the use of these specific safety metrics, the measure provides that a service provider may alternatively use 'equivalent' metrics if appropriate. Equivalent metrics should have the same explanatory value as those we have recommended. If a service provider opts to collect alternative metrics, it should enable the provider to understand whether a recommender system design adjustment affects the risk of users encountering illegal content, compared with the existing variant of the recommender system.

### Incorporating illegal content proxies into this measure

7.25 The safety metrics recommended as part of this measure take account both of illegal content and what we have called 'illegal content proxy'. We recognise that service providers may choose to run their complaints process in a way that does not distinguish between illegal content and content that breaches their terms of service. In chapter 2 of this Volume: 'Content moderation', we recommend that service providers have a content moderation function designed so that, in response to a complaint from a UK user, they can review the content which is suspected to be illegal and either:

- make an illegal content judgement in relation to the content and where the provider determines that it is illegal, swiftly take the content down; or
- where the provider is satisfied that its terms of service prohibit the types of illegal content which it has reason to suspect exist, consider whether the content is in breach of those terms of service, and if it determines that it is, swiftly take the content down.

7.26 We therefore consider that all complaints pertaining to a terms of service breach can be considered as illegal content proxies, provided that those terms of service are sufficiently inclusive of the kinds of illegal harms that are relevant to this measure (see paragraph 7.79).

7.27 Google highlighted the impracticability of isolating UK user complaints about Act-specific breaches from much larger datasets related to terms of service breaches.[1137] We recognise this challenge and consider it to be sufficiently addressed in our measure. As explained in our November 2023 Consultation, complaints data derived from non-UK users included in the on-platform test may be treated as an appropriate proxy for illegal content. This is because the categories of prohibited content within most providers' terms of service do not vary significantly across the jurisdictions in which they operate. If a service provider is satisfied that the categories of content prohibited by its terms sufficiently cover Act breaches when handling UK user complaints (in line with paragraph 7.25), non-UK user

[1136] As defined in paragraphs 7.26 to 7.28

[1137] Email from Google dated 15 July 2024.

complaints resolved under those same categories during a test may also be used for the purposes of generating safety metrics.

7.28 We recognise that using data relating to broader terms of service breaches (including both UK and non-UK complaints) may lead to safety metrics being derived from content that is in breach of a service provider’s terms but does not necessarily amount to illegal content as defined by reference to the types of offences identified in the Act. However, we believe that this data will give service providers an indication of how design adjustments may contribute to the unintentional amplification of illegal content (and thus the risk of users organically encountering that content). We consider this approach preferable to prescribing sampling requirements for a certain level of UK user representation in on-platform testing. This would be a considerably more onerous means of achieving the aims of the measure. Doing so would likely increase the costs of current practice and potentially limit the amount of data available to produce safety metrics.

### What is considered a design adjustment?

7.29 In our November 2023 Consultation, we proposed that providers should produce and analyse safety metrics when conducting on-platform testing of an actual or proposed recommender system design change.

7.30 Our policy intention is that this measure applies to iterative and incremental changes to the design of a recommender system that a provider decides to test. As outlined above, we do not recommend that our measure applies in respect of ‘significant changes’ to the design of recommender systems and our definition of this term for the purposes of Codes expressly carves this out.

7.31 However, we acknowledge that more specific language may be needed to draw out a distinction between significant changes and smaller, more frequent changes made to recommender systems. To help provide clarity for service providers, we have replaced the term ‘design change’ with ‘design adjustment’ in the Code measure.

7.32 In this context ‘design adjustment’ refers to iterative and incremental changes (which we have clarified from ‘small and incremental’) made either to the recommender system’s underlying model(s) or to the algorithms that are responsible for content ranking. A design adjustment may involve any of the following actions (note these are examples and not a definitive list):

- **Expanding or constraining the content pool** – This involves altering the recommender system so that it changes the range of content processed and recommended to users (for example, altering a recommender system so that it analyses content from all accounts, not just those followed by a user). This may be used to increase content diversity or enhance personalisation.

- **Adjusting user and content signals or features** – This involves adjusting the types of cues and signals the system considers when learning about users’ preferences and ranking content accordingly (for example, altering a system so that it determines the relative ranking of a piece of content by analysing how much the content has been liked, viewed, and reposted).

- **Fine-tuning prediction factors** – This involves changing the emphasis that the system places on different predictions made by the model (for example, altering a system so it places greater weight on how likely a user is to comment on a piece of content versus how likely a user is to share that content).

7.33 By contrast, and as set out in our Risk Assessment Guidance there may be other changes made to the design of a recommender system that would amount to a 'significant change' for the purposes of the risk assessment duty and which we do not include within the scope of this measure.[1138] In our Risk Assessment Guidance, we explain that significant changes to the design of a recommender system may include substantial modifications to the recommender system's architecture, algorithms, or objectives; examples might include altering the system's goal criteria, adding new user features, implementing a new ML model, replacing the entire system, or adding a new system. We understand that service providers typically engage in extensive product testing and evaluation of these changes due to their risk of harm and their potential effect on user experience.

7.34 We recognise that on-platform tests are run on design adjustments made to 'live' recommender systems that have already been deployed, and that on-platform tests may therefore not be suitable for significant changes which are likely to undergo other types of testing (i.e., offline tests) prior to deployment. However, we understand that services would be likely to resume on-platform testing as soon as a proposed significant change has completed a suitable and sufficient risk assessment and is deployed as a 'live' system on the service (see the Risk Assessment Guidance). When services decide to resume on-platform testing following the deployment of a significant change, we would expect them to resume the collection of the safety metrics (or equivalent) as per this measure.

7.35 We received feedback from LinkedIn, explaining that service providers will often make changes to their systems that are not always deployed, and that safety metrics should only be collected periodically.[1139] We recognise that not all design adjustments might be deployed for several reasons. For example, poor overall performance and not improving the content discovery experience in a meaningful way. However, as the safety metrics produced allow providers to understand the volume and spread of illegal content and illegal content proxies identified during the testing period for each variant tested, we believe that collecting safety metrics and recording the outcomes is just as valuable for adjustments that are not deployed as it still helps platforms understand the trade-offs and unintended consequences that may occur as a result of design adjustments, and these learnings may contribute to safer design choices in the context of future adjustments.

7.36 However, to avoid creating an undue compliance burden for service providers, safety metrics do not need to be collected in the following certain circumstances:

- changes that amount to a 'significant change' (see paragraphs 7.33 and 7.34) and trigger the risk assessment duty under section 9(4) of the Act;
- changes made in connection with a live and time-sensitive response to a national security risk or other emergency; or
- changes that are not deployed for UK users of the service.

## Benefits and effectiveness

7.37 Where illegal content is shared on a U2U service and is yet to be detected and taken down, recommender systems may play a role in disseminating that content to users. The specific risk we are concerned with here is that recommender systems relevant to this measure may

[1138] Risk Assessment Guidance (Part 3: supporting documents, section titled 'Making a significant change to your service').

[1139] LinkedIn response to November 2023 Consultation, p.14.

increase the likelihood of users being exposed to illegal content. One way this may happen is when a service makes an amendment to its recommender system without appropriately testing the impact on the spread of illegal content.[1140]

7.38 Several studies and journalistic reports highlight the role of recommender systems in the dissemination of illegal content and other harmful content. A working group review from the Global Internet Forum to Counter Terrorism highlighted consensus among experts in the technology, government, civil society, and academic sectors that supports this proposition.[1141] However, the Cyber Threats Research Centre at University of Swansea's response highlighted that there is not sufficient research that suggests the widespread proliferation of illegal content by recommender systems.[1142] They cite several academic papers that investigated this, with the vast majority looking into how recommender systems amplify legal but harmful and borderline illegal content.

7.39 While we recognise that the focus of these studies tends to be on harmful content, we consider that illegal content that is yet to be detected and removed is likely to be disseminated in a similar way. Evidence provided in our Register demonstrates that, whilst recommender systems can confer important benefits, in certain circumstances they can play a role in the dissemination of harmful content, which could indicate an illegal content risk.[1143] This included a systemic review and a study determining that recommender systems can facilitate pathways towards extremist content and content featuring-partially clothed minors.[1144] This evidence suggests that design choices within recommender systems may be an important factor affecting the risk of encountering illegal content as well as harmful content as both types of content are algorithmically disseminated in a similar way.[1145] Evidence in our Register also acknowledges that content recommender systems may increase the risk of certain types of illegal content appearing on a service.[1146] For example, during a time of heightened political activity, the volume of user-generated hate and terror content uploaded onto a service could increase, and the user engagement with that content may also increase (e.g., likes and reshares). This can lead to recommender systems accelerating the spread of hate and terror content.

7.40 Ofcom-commissioned research indicates that it is common for U2U service providers to make frequent adjustments to the design of their recommender systems, with some service

[1140] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content [accessed 8 October 2024].

[1141] Global Internet Forum to Counter Terrorism 2021. Content-Sharing Algorithms, Processes, and Positive Interventions Working Group: Part 1. [accessed 8 October 2024].

[1142] Cyber Threats Research Centre, Swansea University response to November 2023 Consultation, p.12.

[1143] For an overview of the role recommender systems play in increasing risks see the Register of Risks chapter 'Governance, systems and processes'. See also individual harms chapters in the Register of Risks for information regarding the link between recommender systems and each particular kind of illegal harm.

[1144] Cook, J. and Murdock, S., 2020. YouTube is a Pedophile's Paradise. Huffington Post, 20 March. [accessed 8 October 2024]; Whittaker, J., Looney, S., Reed, A., and Votta, F., 2021. Recommender systems and the amplification of extremist content, *Internet Policy Review*, 10 (2). [accessed 8 October 2024]; Yesilada, M., and Lewandowsky, S., 2022. Systematic review: YouTube recommendations and problematic content *Internet Policy Review,* 11 (1). [accessed 8 October 2024].

[1145] Different platforms are likely to have varying volumes of illegal content present at any given period, and this (as well as design choices) may be an influencing factor in the extent to which their recommender system may disseminate such content.

[1146] See Register chapters titled 'Introduction' and 'Governance, systems and processes' providers making hundreds of small changes every week.[1147] Our concern is that some service providers may implement these small changes without fully considering the illegal content risk to users because the risk assessment duty in the Act is only triggered by a 'significant change'.[1148] Our evidence indicates that the incorporation of user reports and moderation decisions (i.e., complaints being upheld) into on-platform tests is one of the most effective ways of identifying unintentional amplification of illegal content.[1149] By observing and tracking these metrics, service providers can uncover emerging patterns of risk exhibited by their recommender systems. This will allow them to improve their understanding of the likely consequences of ongoing design adjustments and make informed decisions regarding future adjustments.

7.41 Ofcom's evidence indicates that on-platform testing can also be an effective way for service providers to observe and respond to patterns of harm through the gathering of safety metrics. It enables providers to identify design features or properties that might be contributing to the dissemination of illegal content (and other types of legal but harmful content).[1150]

7.42 Additionally, the measure will also enable service providers to respond to a changing threat environment in real time.[1151] By observing real-time feedback from users in the form of safety metrics, service providers will be able to uncover emerging and previously hidden patterns of harm. This knowledge will help providers adjust their recommender systems to reduce the unintentional amplification of illegal content or illegal content proxies, thereby reducing the risk of harm to users. Given the impact the design of recommender systems can have on the dissemination of illegal content, the benefits of measures which help improve their design are likely to be material over time. Our conclusion is based on a wide variety of sources, including research, industry expertise, and stakeholder feedback on our original proposal.[1152]

### Benefits of collecting safety metrics

7.43 As set out in paragraph 7.23, this measure recommends that service providers should produce and analyse two safety metrics (or equivalent) that are intended to give service providers an understanding of whether a design adjustment would increase the risk of users encountering illegal content. In this section, we analyse the benefits of these metrics.

- **Metric 1: Total number of items identified as illegal content or as an illegal content proxy** – This metric shows the service provider how many user complaints were upheld as illegal content or as an illegal content proxy. It shows how much of this content was encountered by users in the control and treatment groups during on-platform testing. The metric indicates the overall scale of risk to users in terms of the number of unique

[1147] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content. [accessed 8 October 2024].
[1148] Ofcom 2023. Illegal content risk assessment guidance. [accessed 8 October 2024].
[1149] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content. [accessed 8 October 2024].
[1150] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content. [accessed 8 October 2024].
[1151] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content. [accessed 8 October 2024].
[1152] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content. [accessed 8 October 2024]; Meeting with Rumman Chowdhury, 20 February 2023.

items of illegal content or appropriate proxies contained in the source pool that are surfaced by each variant of the recommender system.

- **Metric 2: Total number of impressions and reach per item of content identified as illegal content or as an illegal content proxy** – The first metric can be used to produce a second metric that indicates the level of user exposure to illegal content across control and treatment groups in terms of impressions and reach per item of content. Impression data is important because it reveals how many times the illegal content or appropriate proxy was encountered. Reach data is important because it shows how many unique user accounts encounter illegal content or appropriate proxies.

7.44 We consider these metrics to be relevant when assessing the risk of users encountering illegal content because they show the distribution of illegal content across a service provider's user base. For example, they can indicate whether a piece of content was recommended multiple times to a limited number of users (high impression, low reach) or whether it was widely distributed to many users but only viewed a few times (high reach, low impression).

7.45 Meta and Google highlighted the importance of allowing service providers flexibility in how they approach safety testing of recommender systems.[1153] While we recognise that service providers may have alternative and effective methods for evaluating their systems, our proposal is informed by the good practice identified in our commissioned research.[1154] The measure still allows for a degree of flexibility in many areas. For example, the measure does not prescribe which type of on-platform test should be deployed.

7.46 As outlined at paragraph 7.24 above, the measure also gives service providers the flexibility to establish their own safety metrics, which are equivalent to those we have recommended in the sense that they enable the provider to understand whether a design adjustment would increase the risk of users encountering illegal content. If a provider does so, we would expect those equivalent metrics to have a similar explanatory value and therefore be an effective tool to monitor risk.

7.47 The metrics would also enable the provider to run a comparative analysis across all variants of the recommender system tested to evaluate the respective illegal content risk.

### Benefits and effectiveness of on-platform tests and log of test results

7.48 This measure recommends the collection of these metrics when conducting on-platform tests (such as A/B tests) of design adjustments to their recommender systems. However, there are many ways of evaluating the effects of a recommender system on users, including user surveys, sock puppet accounts,[1155] and debugging exercises.[1156] Unlike some other

[1153] Google response to November 2023 Consultation, pp.57-58; Meta response to November 2023 Consultation, p.32.

[1154] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content. [accessed 8 October 2024].

[1155] In the context of recommender systems evaluation, 'sock puppets' are artificial user accounts or bots created to simulate real user behaviour for experimentation purposes. Sock puppet accounts are often set up by researchers seeking to observe how a recommender system serves content to accounts exhibiting different behaviours and preferences.

[1156] 'Debugging' refers to the software engineering process of tracing the exact cause of an anomaly (such as a spike in illegal content or harmful content recommendations). This is a resource-intensive process that involves specialist staff methodically uncovering the internal computational process that caused a particular anomaly or incident.

evaluation techniques, on-platform tests can allow for direct causal inferences to be drawn, if they produce statistically significant patterns. On-platform tests take the form of randomised controlled trials, which are widely regarded as the gold standard of research to establish causal effects. This class of testing was identified as one of the most robust methods for evaluating recommender systems in research commissioned by Ofcom.[1157]

7.49 We have also considered other available evidence regarding the effectiveness of conducting on-platform tests and keeping logs, recording those test results. In 2020, LinkedIn revealed that it had established inequality metrics for monitoring barriers to economic opportunity for its users (in the form of exposure to job notifications). These inequality metrics were then monitored during on-platform tests of product changes (including for LinkedIn's recommender system).[1158] In 2021, X (formerly Twitter) released details of a large A/B/x test which examined how a proposed change to its recommender system altered the dissemination of political content in user feeds.[1159]

7.50 This measure recommends the use of logs to record the results of on-platform tests, and other relevant information such as key design features and decisions taken. Logs, which are commonly used in the management and testing of algorithmic systems, provide a structured and historic account of how a recommender system responds to design adjustments. This allows product and engineering teams to continuously monitor changes in risk patterns over time and identify those design adjustments that increase risks to users.

7.51 Evidence obtained through discussion with Rumman Chowdhury, an expert on recommender systems with extensive experience in algorithmic governance and on-platform tests, indicates that it is normal practice for service providers to maintain logs that record test results which explain the impact of product updates (or design adjustments). These logs detail the performance of recommender systems according to commercial metrics.[1160] Rumman Chowdhury indicated that relevant teams typically have access to these results and that they are used to inform product changes (in this case, design adjustments) as appropriate. We conclude that records of safety metrics as proposed by this measure are likely to be referred to by these same teams, making a log that records those results an effective resource in the ongoing management of risk associated with recommender systems.

### Using the metrics as a means of identifying and managing risk to users

7.52 The measure asks service providers to refer to the log recording on-platform test results (including the safety metrics obtained) when making future design adjustments. However, the measure does not go so far as to recommend that service providers should act on the results of every single test or opt for a design adjustment that appears to be the safest for users based on a single test result.

[1157] Ofcom, 2023. Evaluating recommender systems in relation to illegal and harmful content [accessed 8 October 2024].

[1158] LinkedIn Engineering (Saint-Jacques, G., Sepehri, A., Li, N. and Perisic, I.) 2020. Building inclusive products though A/B testing. [accessed 8 October 2024].

[1159] Global Partnership on Artificial Intelligence (GPAI), 2022. Transparency Mechanisms for Social Media Recommender Algorithms. [accessed 8 October 2024].

[1160] Meeting with Rumman Chowdhury, Monday 20 February 2023.

7.53 As mentioned in paragraph 7.7, some stakeholders have expressed concern that this limits the impact of the measure, permitting service providers to disregard test results.[1161]

7.54 This measure is based on the practice of continuous monitoring in software and AI governance. This is the on-going process of observing the outputs of an algorithmic system to ensure that it is operating as intended. When used in the context of recommender systems, this practice can help service providers identify unintended patterns of content dissemination and, over time, be able to determine what design adjustments might be affecting the risk of users encountering illegal content. As a result, the benefits of this measure will be realised in the long term as service providers increase their risk awareness of how different design adjustments affect the risk of illegal content dissemination. While this measure requires service providers to ensure that the log of the results of previous on-platform tests in the context of future design adjustments, we have not included a specific requirement to act upon any set of test results in the measure. We do not consider it proportionate at this stage to include within the measure a recommendation that service providers act upon every individual test result in a particular way. There are two reasons for this:

- We recognise that each test may not produce a set of results that are meaningful or statistically significant. Providers need to conduct multiple tests before they can build a reliable picture of the impact of design adjustments.
- The safety metrics that we are recommending be observed in this measure only relate to illegal content (or where relevant, illegal content proxies). Services may also want to consider how the design of their recommender systems alters user exposure to other types of harmful content, as well as the overall performance of the system across a variety of metrics. We want to avoid a situation where services make a design choice that reduces user exposure to illegal content in order to adhere to this measure while increasing their exposure to other types of harm.

7.55 While the measure does not involve service providers taking specific actions after every test result, it does establish several steps to secure that the test results will influence future design choices. This includes a requirement for services to maintain a log that records the test results, to explain the design decision taken following a test, to make that log available to relevant internal teams, and to have those teams refer to that log when making future adjustments to the recommender system. Service providers should ensure that relevant staff refer to the log in the context of future recommender system design adjustments. Service providers may be expected to demonstrate that they have done this. Therefore, we are confident that this measure creates increased accountability for service providers which, over time, will improve the safety of design choices made in relation to recommender systems compared to a counterfactual where service providers are not producing and analysing safety metrics.

[1161] Integrity Institute response to November 2023 Consultation, pp.17-19; Molly Rose Foundation response to November 2023 Consultation, pp.32-33; NSPCC response to November 2023 Consultation, pp.39-40.

## Costs and risks

### Costs

7.56 In our November 2023 Consultation, we considered the one-off and ongoing costs of this measure for service providers that already run on-platform tests, and for whom this measure is recommended. Such service providers will already have an established testing environment in place, along with the specialist staff needed to execute on-platform tests and implement the recommendations put forward in this measure. We considered several additional costs in this scenario.

- **One-off cost** – Setting up new safety metrics (see paragraph 7.23) or equivalent metrics will require service providers to identify the relevant data points, establish a data cleaning and preparation process, and establish a formula for analysing that data to produce the safety metrics. This may require time from in-house data engineers and data scientists.
- **Ongoing cost** – In addition to the one-off cost, we expect service providers to incur additional ongoing costs related to the maintenance of the measure to ensure it continues to perform as intended. Service providers will also need to collect and store additional data for the duration of the tests they perform. They will need to store data relating to all pieces of content shown to users in the treatment and control groups, including information about the content's classification (deemed illegal or otherwise), number of impressions, and reach. They will also need to maintain a log of past test results. There may also be ongoing costs of an extended product management cycle. Additional staff time may be needed to review the new metrics produced as part of this measure and decide on how to act on them. This may require additional time commitment from in-house data engineers and product teams.

7.57 We had limited feedback in response to the November 2023 Consultation on the specific direct costs that are needed to implement this measure and have kept the cost assumptions and estimates largely unchanged.[1162]

7.58 Regarding the one-off cost of designing and setting up of the new safety metrics, Rumman Chowdhury explained that a new safety metric she helped establish required 2,000 human hours split equally between software engineering time and research time.[1163] This amounts to an approximate one-off set-up cost of £70,000 to £140,000.[1164]

7.59 Our analysis indicates that the largest U2U services that already perform on-platform tests would be able to meet these one-off costs of this measure. We also consider the same is likely to be true of smaller services that run on-platform tests. For a smaller service to run on-platform tests already, they would have needed to invest a significant amount in testing infrastructure, which would indicate that they can afford the moderate upfront cost of creating new safety metrics.

7.60 We consider the cost of establishing the safety metrics set out in this measure likely to be towards the lower end of (or potentially below) the range suggested. As discussed in

[1162] We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by the Office of National Statistics (ONS). We received some feedback on the general cost assumptions (such as salary assumptions) that are fed into these costs. We consider that feedback in Annex 5

[1163] Meeting with Rumman Chowdhury, 20 February 2023.

[1164] We estimate the costs assuming 1000 hours for engineering time and an equivalent time input from professional occupation staff (researcher) using the wage assumptions as set out in Annex 5.

paragraph 7.27, we assume that many service providers already capture much of the required data (such as complaints data). Therefore, many providers will already be in alignment with our expectations for the construction of the metrics outlined in this measure. Service providers will also have the flexibility to set up their own relevant signals enabling them to observe illegal content risk on an ongoing basis, if those metrics are equivalent in their explanatory value.

7.61 Although not directly relevant to this measure, we are aware that some larger service providers have prior experience in developing other types of safety metrics. For example, YouTube produces a quarterly metric known as the 'violative view rate'. This is based on reviewing a random sample of videos and identifying those which breach YouTube's Community Guidelines.[1165] Meta produces a metric known as 'prevalence' which allows it to estimate the percentage of total views of content that breach its Community Standards across Facebook and Instagram.[1166]

7.62 In addition to this one-off cost, we assume an annual ongoing cost of maintaining the safety metrics as part of regular updates. This is estimated to be approximately 25% of the one-off costs, at around £17,500 to £35,000.[1167]

7.63 Rumman Chowdhury expressed the view that the ongoing cost of storing the metric data is likely to be negligible (considering the limited additional data that would require storage). This is especially true for the largest U2U service providers as they already operate large data storage centres. Data storage costs are likely to be further limited by the fact that this data will not need to be retained beyond the duration of tests (which we understand do not typically run for more than several weeks at a time). The additional expense of storing the results log would be minimal since it contains only aggregate high-level information, and we understand that logs are likely to be an integral and existing component of a provider's existing testing infrastructure.

7.64 Regarding the ongoing costs, we do not consider this is likely to amount to a disproportionate expense for those services that already run on-platform tests – even if services perform upwards of hundreds of tests per week. This is because services would already be dedicating resources to reviewing the other metrics being measured through tests, and thus this measure only extends an existing exercise rather than creating a new one. Moreover, this measure requires that only two additional metrics be observed and analysed, and does not specify the nature of that analysis, which services are free to perform as they choose and in a manner that is efficient to them.

7.65 The effect of these costs may be lessened to some extent by addressing potential causes of harm upfront, and thereby reducing the costs a service provider incurs in mitigating harm after the fact. For example, reducing the extent to which recommender algorithms disseminate illegal content may reduce the costs incurred by content moderation teams when dealing with reports of illegal content.

7.66 Google suggested that the measure risks creating a significant compliance burden for providers that have on-platform testing.[1168] Another issue raised by Google was that the

[1165] YouTube (O'Connor, J.), 2021. Building greater transparency and accountability with the Violative View Rate. [accessed 8 October 2023].
[1166] Meta, 2022. Prevalence. [accessed 8 October 2024].
[1167] Based on our standard assumptions set out in Annex 5.
[1168] Google response to November 2023 Consultation, pp.57-58.

threshold for what is considered a design adjustment that triggers the collection of safety metrics was too low, making the measure disproportionate.[1169] As explained in the 'How this measure works' section, this measure does not recommend that service providers conduct testing procedures in addition to what they already do. Moreover, we have considered the compliance burden in the design of the measure – for instance, in paragraphs 7.27 and 7.28 where we provide flexibility to services on collecting equivalent metrics and in paragraph 7.36 where we explain circumstances where the recommended metrics may not need to be collected. However, if a service provider decides that a design adjustment is sufficiently material or non-trivial to warrant conducting an on-platform test to evaluate its impact on engagement metrics, then that adjustment also warrants the collection of safety metrics as part of those evaluations.

7.67 We received subsequent and clarificatory feedback from Google that this measure may detract from meaningful compliance due to similar testing requirements that may be required under other regulatory regimes. In the 'Our approach to developing Codes measures' chapter, we have recognised that there may be challenges for providers captured by multiple regulatory regimes, and we have sought to reduce the regulatory burden on providers where possible. Our measure is flexible insofar as it provides providers with the option of collecting alternative, equivalent safety metrics. It also remains open to providers to employ alternative measures to comply with the safety duties in the Act.[1170]

### Risks

7.68 We recognise that there may be ethical concerns associated with the use of on-platform tests, as users in treatment groups may be exposed to more illegal content than they would otherwise encounter. This measure will not increase these risks because it does not specify a recommendation for providers to perform any new on-platform tests, but rather specifies only that service providers collect additional metrics within existing on-platform tests. We maintain that these considerations do not render the measure unethical or ineffective.

7.69 Separately, Snap, Meta, and Google said that the measure might discourage on-platform testing altogether.[1171] We recognise this risk, but conclude this to be low, not least because of the enormous value that providers derive from using on-platform tests to monitor changes in commercial metrics (such as engagement scores like clicks and views).

## Rights impact

### Freedom of expression

7.70 As explained in 'Introduction, our duties, and navigating the Statement', as well as in chapter 14 of this Volume: 'Statutory tests', Article 10 of the ECHR sets out 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. Ofcom must not interfere with this right unless it is satisfied that it is prescribed by law, corresponds to a pressing social need and is proportionate to the legitimate aim pursued.

7.71 This measure does not have a direct impact on the right to freedom of expression as it focuses on generating organisational risk awareness from which safety conscious design

[1169] Email from dated 15 July 2024.

[1170] Letter from Google dated 18 November 2024.

[1171] Google response to November 2023 Consultation, pp. 57-58; Meta response to November 2023 Consultation, p.32; Snap response to November 2023 Consultation, pp.21-22.

decisions are made. There may be some indirect effect to the extent that, as a consequence of collecting the recommended safety metrics, service providers design their recommender systems so as to avoid recommending illegal content. This may also have an impact on some content that is not illegal (but which may be contrary to a service's terms of service). However, the content will still be present on the service. To the extent that there was any indirect impact on users and service's right to freedom of expression, we would consider it to be proportionate in pursuit of a legitimate aim since it would reduce users' exposure to illegal content.

### Privacy

7.72 As explained in 'Introduction, our duties, and navigating the Statement', as well as in chapter 14 of this Volume: 'Statutory tests', Article 8 of the ECHR sets out the right to respect for private and family life. An interference with this right must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest. In order to be 'necessary', the restriction must correspond to a pressing social need, and be proportionate to the legitimate aim pursued.

7.73 Service providers have a duty to operate a complaints procedure that allows users and affected persons to complain about content which they consider to be illegal content, sets out the action we consider appropriate to take.[1172] This means that we expect providers to be determining complaints in any event, so we do not consider that the need to consider complaints for the purposes of this measure should necessarily give rise to any additional privacy implications. Complaints data could be anonymised before processing for the purposes of this measure.

7.74 While the collection of safety metrics (including the number of impressions and reach per item of illegal content identified) involves monitoring of what users in each of the treatment and control groups see, the measure recommends that the safety metrics be obtained only when other tests are being carried out which are also likely to involve a significant degree of monitoring of their activities. As our measure requires an additional form of monitoring within these tests, it may to some extent amount to an interference with user privacy. However, we regard it as proportionate in order to achieve the legitimate aim of protecting users from the harm caused by illegal content because, by implementing this measure, service providers will be able to operate their content recommender systems in a way that minimises the risk of users organically encountering illegal content.

### Data protection

7.75 We recognise that this measure will involve the processing of users' personal data in various aspects of its implementation, as outlined below. As such, service providers will need to comply with data protection legislation and should refer to relevant guidance from the ICO.[1173]

7.76 As set out in paragraph 7.7, ICO highlighted that the collection of the specified safety metrics might also entail the processing of personal data and in particular, that providers adopting this measure should ensure that they comply with the purpose limitation and data minimisation principles.[1174] [1175] In relation to the metric relating to the total number of

[1172] See chapter 6 in this Volume: 'Reporting and complaints' for further information.

[1173] ICO, UK GDPR guidance and resources.

[1174] ICO response to November 2023 Consultation, p.20.

[1175] ICO, A guide to the data protection principles.

impressions and reach per item of content identified as illegal content across groups, we recognise that this may involve the processing of new personal data or further processing of existing personal data. The safety metrics also involve the consideration of the total number of items identified as illegal content or an illegal content proxy in response to complaints during the testing window. This will not be an additional processing of personal data since in our view it would need to happen in any case, and it is not clear that a service provider would need to process additional personal data in order to consider this metric. However, to the extent they did so they would need to ensure compliance with applicable data protection legislation.

7.77 On-platform testing also involves the allocation of users into different groups, as explained at paragraph 7.11. We therefore acknowledge that this aspect of implementation may also have a potential impact on user's rights under data protection legislation. Allocation is likely to involve processing of personal data, and all subsequent monitoring of what content users see and how they react to it would be processing of personal data. However, we note that our measure will only apply to providers who are already carrying out on-platform testing, which means this impact is likely to be fairly small, since much of the processing would happen in any event.

7.78 Overall, we consider that the impact of this measure is likely to be limited where providers comply with relevant laws (as outlined above), and that any interference is both necessary to ensure compliance with the Act and proportionate.

## Who this measure applies to

7.79 In our November 2023 Consultation, we recommended the measure should apply to services where providers have assessed as being at medium or high-risk for at least two kinds of specified illegal harms in their latest illegal content risk assessment: terrorism; child sexual abuse material ('CSAM') (CSAM URLs or image-based CSAM); encouraging or assisting suicide[1176]; harassment, stalking, threats and abuse; hate offences; drugs and psychoactive substances; extreme pornography offences; intimate image abuse offences; foreign interference offences.[1177] We also recommended that the measure should only apply in relation to services that already run on-platform tests.

7.80 Some stakeholders expressed concern about the limits of who the measure would apply to. The NSPCC said that the measure should apply to providers of all large multi-risk services regardless of whether they already have on-platform testing.[1178] While we recognise the appetite for extending the measure in this way, we must ensure that any requirements placed on service providers are proportionate to the expected safety benefits. Based on our understanding of the costs of establishing new testing infrastructure, which are significant, we are not at this point persuaded that this would be proportionate. That said, we understand that most providers that operate recommender systems would very likely be

[1176] The measure as consulted on in the November 2023 Consultation applied to services which are at medium or high risk of 'Encouraging or assisting suicide' and 'Encouraging or assisting serious self-harm'. We have made sure the harms groupings only include priority offences, consistent with Parliament's decision that they should be a priority. Consequently, we have taken 'Encouraging or assisting self-harm' out of scope.

[1177] See Register of Risks chapters titled 'Terrorism'; CSEA (specifically section on Child sexual abuse material (CSAM)); 'Hate'; 'Harassment, stalking, threats, and abuse'; 'Intimate image abuse'; 'Extreme pornography'; 'Drugs and psychoactive substances'; 'Encouraging or assisting suicide'; 'Foreign interference; 'Non-priority offence - Encouraging or assisting self-harm'.

[1178] NSPCC response to November 2023 Consultation, pp.39-40.

running tests on those systems already, and so the number of providers in scope of the measure would already be sizeable.

7.81 While not specific to this measure, we received responses from stakeholders on our approach to using multi-risk criteria when recommending who a measure applies to as opposed to applying the measure equally to services identifying a single-risk of harm.[1179] The case for extending the measure so it applies to all single-risk service providers is more finely balanced. As set out in 'Our approach to developing Codes measures', we intend to consider the case for extending this measure to single-risk services in a future consultation in Spring 2025.

7.82 Overall, we have decided that the measure applies to all U2U service providers that:

- already employ on-platform testing of their recommender systems; and
- have assessed that their service is at high or medium risk for two or more types of the illegal harms identified in paragraph 7.79, as identified in their latest illegal content risk assessment.

## Conclusion

7.83 Our analysis shows that the measure we are recommending is likely to deliver material benefits and that the costs and impacts on rights that will result from it are modest and proportionate. We have therefore decided to leave the measure unchanged from the wording proposed in our November 2023 Consultation, with the exception of replacing the term 'design change' with 'design adjustment'.

7.84 When running on-platform tests on design adjustments to their recommender systems, all U2U service providers should collect the additional safety metrics (or equivalent metrics that hold similar explanatory value) recommended as part of this measure. Service providers should also maintain a log of these metrics, detailing the impact of the design adjustment and an explanation of why a particular adjustment was deployed. This log should also be made available to relevant staff working on recommender systems. By implementing this measure, we maintain that services will be able to operate their recommender systems in a way that enables them to monitor, identify, and mitigate illegal content risk in real time.

7.85 The full text of the measure can be found in our Illegal Content Codes of Practice for User-to-User services, in which it is referred to as ICU E1. This measure is part of our Codes on terrorism, child sexual exploitation and abuse ('CSEA') and other duties.

[1179] Our approach to developing Codes measures.

# 8. U2U settings, functionalities and user support

## What is this chapter about?

This chapter describes a series of measures we are recommending to tackle online grooming, why we are recommending them, and to which providers of user-user (U2U) services they should apply.

## What decisions have we made?

We are recommending the following measures (The measures detailed below apply in relation to users aged under 18):

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| ICU F1 | Providers should implement **safety defaults** for child user accounts, which target particular **functionalities**[1180] and that restrict the visibility and engagement between **child users** and other **users**. | Providers of U2U services which have an **existing means to determine the age or age range of a particular user** of the service and have relevant functionalities, if they are:<br>• at high risk of grooming, or;<br>• are large services and at medium risk of grooming. |
| ICU F2 | Providers should give **child users** relevant **supportive information** at critical points[1181] in a child user's journey to allow child users to make more **informed choices**. Providers should ensure the messaging is prominently displayed, and is clear and easy for children to understand. | Providers of U2U services which have an **existing means to determine the age or age range of a particular user** of the service and have relevant functionalities, if they are:<br>• at high risk of grooming, or;<br>• are large services and at medium risk of grooming. |

## Why have we made these decisions?

Child sexual exploitation and abuse (CSEA) is a serious crime which can have a severe and lifelong impact on children and communities. Grooming for the purpose of sexual abuse involves a perpetrator establishing communications with a child to enable their abuse and exploitation both online and offline.  In online spaces, it can often lead to the exchange of self-generated indecent images and financially motivated sexual extortion, which would also

[1180] Measure ICU F1 will target functionalities which have been identified as risk factors for grooming harm on U2U services. These include network expansion prompts, direct messaging, connection lists and automated information displays.

[1181] As part of measure ICU F2, we specify four critical points in a child user's online journey, where the child user may take a decision which impacts their engagement with a user or users. See 'Measure on support for child users' in volume 2: chapter 8: U2U settings, functionalities, and user support, for more details.

mean that a range of other specific child sexual abuse material (CSAM) offences have been committed.

Strategies that perpetrators deploy to groom children frequently include: sending scattergun 'friend' requests to large volumes of children; infiltrating the online friendship groups of children they have succeeded in connecting with; and sending unsolicited direct messages to children they are not connected with. Taken together, the measures described in this chapter will make it more difficult for perpetrators to adopt these strategies and empower child users to make informed choices about their online interactions. This would therefore make grooming more difficult, thereby combating CSEA.

## Introduction

8.1 The measures recommended in this chapter are primarily designed to combat grooming for the purposes of child sexual exploitation and abuse ('CSEA') on user-to-user ('U2U') services.[1182] As explained in our November 2023 Illegal Harms Consultation ('November 2023 Consultation'), grooming involves a perpetrator communicating with a child with the intention of sexually abusing them either online or in person.[1183] This preparatory grooming behaviour (which can be in and of itself an offence) can be carried out anywhere. In online spaces, it can often lead to the exchange of self-generated indecent images and financially motivated sexual extortion, which would also mean that a range of other specific child sexual abuse material (CSAM) offences have been committed.[1184] Grooming can cause severe and lifelong harm. For a detailed analysis of the risks of grooming online and the severe harm it causes, see the chapter titled 'CSEA' in the Register of Risks ('Register').[1185]

8.2 The recommended measures in this chapter aim to assist providers of U2U services to comply with their duties under the Online Safety Act 2023 ('the Act') to take steps 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.[1186] The measures recommended in this chapter assist with the compliance of particular safety duties, and are relevant to specific areas in the Act under which providers are required to take measures if proportionate to do so.[1187]

8.3 While the measures outlined in this chapter focus primarily on addressing the risk of grooming online, they also have the benefit of addressing risk factors that contribute to other kinds of illegal harms.[1188] We argue these measures can therefore help providers

[1182] Under the Online Safety Act 2023 ('the Act'), the recommended measures in the Codes apply only to the operation of the service in the UK or as it affects UK users.
[1183] Ofcom, 'Protecting people from illegal harms online', Volume 4: How to mitigate the risk of illegal harms – the illegal content Codes of Practice, November 2023, p.231.
[1184] Multiple grooming offences are listed as priority offences in Schedule 6 of the Act. Please refer to Ofcom's Illegal Content Judgements Guidance for more information on the priority offences.
[1185] See the Register of Risks chapter titled 'Child Sexual Exploitation and Abuse (CSEA)' for more details.
[1186] See section 10(2)(a)(b)(c) of the Online Safety Act, 2023, for further details.
[1187] ICU F1 will address 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 (section 10(4)(d)) and functionalities, allowing users to control the content they encounter (section 10(4)(f)). ICU F2 will address user support measures (section 10(4)(g)).
[1188] This includes (but is not necessarily limited to) harassment, stalking, threats and abuse, hate, controlling or coercive behaviour (CCB), or even terrorism.

reduce the risk of harm related to other offences with similar risk factors as grooming. We discuss this further in the 'Benefits' section for both measures.

## Age of children covered by the measures

8.4 In formulating the measures discussed in this chapter, we considered what age threshold would be appropriate – namely, whether the measure should be applied to users under the age of 16 or users under the age of 18. The term 'child' captures a broad range of ages and social and cognitive development stages, and we are conscious that our recommendations must be effective at safeguarding children from online harms while not unduly restricting the online lives of older children, particularly those between 16 and 18 years old.

8.5 As set out in our Register chapter titled 'CSEA', the priority offences grouped under the category of 'grooming offences' feature the shared characteristic of an abuser developing a relationship with a child to facilitate CSEA.[1189] [1190] Many specific priority offences – such as meeting a child following sexual grooming and sexual communication with a child – apply only if a child is under 16 years. However, other grooming offences relating to sexual exploitation, including those relating to the generation of child sexual abuse material (CSAM), relate to children up to the age of 18.[1191] We also have evidence that suggests these offences are committed against children in the 16 to 18 age range.[1192] In our November 2023 Consultation, we therefore considered it would be appropriate to apply the measures to all child users under 18 years.

8.6 Two respondents agreed with our assessment concerning the age of children covered by the measures.[1193] The Children's Commissioner for England "strongly" supported our recommendation that all children under 18 years old should be treated as child users in relation to the safety defaults measure.[1194] We Protect Global Alliance also welcomed both measures applying to all users under the age of 18.[1195]

8.7 Protection Group International however disagreed with our assessment of the measures' applicability to all children under 18 years old. It argued it was "an umbrella approach", with evidence suggesting that older children aged 15-17 disliked being seen or treated as younger children.[1196] It further argued that Ofcom should differentiate between under 18-year-olds and other child age groups.

8.8 We recognise that older children may dislike being treated the same as younger children for the purposes of our measures. However, our evidence indicates that a significant risk of harm from grooming is present for children of all age groups, including older children (who may be at risk of financially motivated sexual extortion, often colloquially known as

[1189] See Register of Risks chapter titled 'CSEA' for more details.

[1190] Schedule 6 of the Online Safety Act 2023.

[1191] Section 15 of the Sexual Offences Act 2003 and article 22 of the Sexual Offences (NI) Order 2008 (S.I. 2008/1769 (N.I. 2)).

[1192] See Register of Risks chapter titled 'CSEA' for more details.

[1193] Children's Commissioner for England response to November 2023 Illegal Harms Consultation, p.23; We Protect Global Alliance response to November 2023 Illegal Harms Consultation, pp.19-20.

[1194] Children's Commissioner for England response to November 2023 Consultation, p.23.

[1195] We Protect Global Alliance response to November 2023 Consultation, pp.19-20.

[1196] Protection Group International response to November 2023 Illegal Harms Consultation, pp.9-10.

'sextortion').[1197] We have therefore recommended that our measures apply to all children under 18 to ensure that older children also have protections against CSEA.

8.9 We also appreciate that our approach could have some impact on children's right to freedom of expression. However, we consider this proportionate given the risk of harm to older children (as described in the 'Rights impact' sections). Furthermore, the measures involve the settings being on by default, meaning the settings can be 'turned off' if a child decides that it does not want them. If they do 'turn off' the settings however, they will be provided with supportive information (as outlined in our second measure), informing them of the risks of doing so.

8.10 Having reviewed the feedback, we have decided to continue with our proposed approach. We consider it appropriate to apply the measures in this chapter to all child users under 18 years, to ensure that 16- and 17-year-olds have these protections.

## Measure on safety defaults for child users

8.11 In our November 2023 Consultation, we proposed that services that have existing means to identify child users and have particular functionalities, should implement the following safety defaults (previously 'default settings') for child user accounts.[1198] [1199]

- Child users should not be presented with network expansion prompts or included in network expansion prompts presented to other users.
- Child users should not be visible in the connection lists of other users. The connection lists of child users should also not be visible to other users.
- Child users should not receive direct messages from users they are not connected to.
- Where the service has no user connection functionality, child users should not receive direct messages from users unless they actively confirm to receive the messages.
- Automatically generated and displayed location information relating to child users' accounts should not be visible to other users via a child user's profile, content posts, or live location functionalities.

8.12 We specified that this measure applies to users aged under 18 and proposed that the measure (i) would only apply to services to the extent that a service has an existing means of identifying child users and (ii) would apply where the information available to services indicated that a user is a child.

8.13 We also proposed that services already using age assurance should use this to determine whether a user is a child for the purposes of the protections under these measures.

8.14 In our November 2023 consultation, we set out that, in due course, we would consult on proposals related to age assurance and that, when any relevant guidance on age assurance came into force, our provisional expectation would be that providers in-scope of this

[1197] See Register of Risks chapter titled 'CSEA' for more details.

[1198] The intended effect of the measure is for the safety defaults to be specifically applied to child user accounts. However, for simplicity in our explanation of the measures, we have referred to 'child users' when describing the effect and implementation of the measure in this chapter.

[1199] For clarity, for the rest of this chapter, we have amended our measure's title wording 'default settings' included in our November 2023 and May 2024 Consultations to 'safety defaults' to better reflect our measure's intended outcome.

measure and the other measure recommended in this chapter (support for child users) would use age assurance, as defined by that guidance, to determine whether a user is a child. We further address this in ‘How this measure works’, specifically under the ‘How this measure will interact with the Children’s Safety Codes’ heading.

8.15 We proposed that this measure should apply to providers of:

- all services at high risk of grooming; and
- all large services at medium risk of grooming.

## Summary of stakeholder feedback[1200]

8.16 Civil society organisations, the Children’s Commissioner for England, the Information Commissioner’s Office (ICO), [✂], an identity service provider and several others, expressed their broad support for this measure.[1201] However, some of the support was caveated, as we explain from 8.19 and subsequent paragraphs.

8.17 In terms of respondents who expressed support, Barnardo’s “welcomed action” that would make it more difficult for adults to groom, abuse, or sexually exploit children.[1202] The National Society for the Prevention of Cruelty to Children (NSPCC) “strongly” supported the safety defaults’ requirements surrounding direct messaging features, and suggested that network expansion prompts are “particularly risky” for children.[1203] INVIVIA also suggested that the measure’s focus on safe default settings for children was “commendable”, and represented a “proactive approach” to children’s online safety.[1204]

8.18 Alongside the November 2023 Consultation, we commissioned Praesidio Safeguarding to carry out an engagement piece with children and young people regarding their views on the safety defaults measure. Broadly, the children we consulted were largely supportive of the measure proposed and felt that they would help increase their safety online. Children noted the default approach as choice preserving, as it allowed them to tailor their experiences while making personal decisions about their safety. Children who had received unsolicited sexual messages from people they did not know were particularly supportive.[1205]

[1200] Note this list is not exhaustive, and further responses can be found in Annex 1: ‘Further stakeholder responses’.

[1201] 5Rights Foundation response to November 2023 Consultation, p.26; Barnardo’s response to November 2023 Illegal Harms Consultation, pp.19-20; Betting and Gaming Council’s response to November 2023 Illegal Harms Consultation, p.10; Children’s Commissioner for England response to November 2023 Consultation, p.22; Duran Dwyer response to November 2023 Illegal Harms Consultation, p.7; Information Commissioner (ICO) response to November 2023 Illegal Harms Consultation, pp.19-20; INVIVIA response to November 2023 Illegal Harms Consultation, pp.19-20; Nexus response to November 2023 Illegal harms Consultation, pp.15-16; NSPCC response to November 2023 Illegal Harms Consultation, p.34; One ID Ltd response to November 2023 Illegal Harms Consultation, pp.2-3; Segregated Payments Limited response to November 2023 Illegal Harms Consultation, p.11; [✂]; UK Safer Internet Centre (UKSIC) response to November 2023 Illegal Harms Consultation, p.12; We Protect Global Alliance response to November 2023 Consultation, pp.19-20.

[1202] Barnardo’s response to November 2023 Consultation, pp.19-20.

[1203] NSPCC response to November 2023 Consultation, pp.35-36.

[1204] INVIVIA response to November 2023 Consultation, pp.19-20.

[1205] Praesidio Safeguarding, 2024. Consulting children on proposed safety measures against online grooming. [accessed 16 December 2024].

8.19 We identified several themes from respondents' feedback related to the measure:[1206]

- age assurance & "identifying" child users;
- feedback on how this measure works;
- feedback on the effectiveness of the safety defaults;
- costs of the measure on users' online experiences; and
- feedback on who the measure applies to.

8.20 We summarise these themes in the following paragraphs.

## Age assurance

8.21 While respondents were generally supportive of the measure, several noted the absence of an age assurance recommendation and some suggested that introducing it could strengthen the effectiveness.[1207] VerifyMy argued that the absence of an age assurance requirement was a "missed opportunity", given the wide range of age estimation technologies available.[1208] INVIVIA said that requiring age assurance would increase the effectiveness of the safety defaults measure, but cautioned that users' privacy should be protected if this was introduced.[1209] NSPCC suggested that not including "age-checking" processes limited the measures' effectiveness.[1210] We Protect Global Alliance also called for a "more comprehensive approach" to the implementation of age assurance technologies on U2U services.[1211]

8.22 Children in our engagement piece also highlighted the need for effective age assurance to prevent children from misrepresenting their ages and missing out on the potential benefits from the recommended safety defaults.[1212]

8.23 We respond to these concerns in the 'How this measure works' section (paragraphs 8.50 to 8.58).

[1206] This list is not exhaustive. We address additional stakeholder feedback in Annex 1: 'Further stakeholder responses'.

[1207] Barnardo's response to November 2023 Consultation, pp.19-20; INVIVIA response to November 2023 Consultation, p.19]; [✂]; NSPCC response to November 2023 Consultation, p.35; NWG response to November 2023 Illegal Harms Consultation, p.9; UKSIC response to November 2023 Consultation, p.40; VerifyMy response to November 2023 Illegal Harms Consultation, p.11; We Protect Global Alliance response to November 2023 Consultation, pp.19-20; Praesidio Safeguarding, 2024. Consulting children on proposed safety measures against online grooming. [accessed 16 December 2024].

[1208] VerifyMy response to November 2023 Consultation, p.11.

[1209] INVIVIA response to November 2023 Consultation, p. 19.

[1210] NSPCC response to November 2023 Consultation, p.35.

[1211] We Protect Global Alliance response to November 2023 Consultation, pp.19-20.

[1212] Praesidio Safeguarding, 2024. Consulting children on proposed safety measures against online grooming. [accessed 16 December 2024].

### Self-declaration

8.24 Stakeholders stated that self-declaration is not an adequate form of age assurance, as children often lie about their age.[1213] [1214] The UK Safer Internet Centre (UKSIC) also noted concerns that the measure could be easily circumvented through self-declaration in the absence of age association requirements.[1215]

8.25 We address this issue in ‘How this measure works’ section (paragraphs 8.50 to 8.58).

### Exclusion of services who cannot “identify” children

8.26 Several civil society organisations, service providers, and the trade body for age verification providers raised concern about the exclusion of services with no existing means of identifying child users from scope of this measure.[1216] Some stakeholders stated it could act as a disincentive, discouraging providers from introducing age assurance.[1217] Specifically:

- Yoti felt it odd to indicate to a service provider that they were not within the scope of the measure if they do not have such means in place.[1218]
- 5Rights Foundation “strongly disagreed” with service providers being considered out of scope if they did not have a means to identify child users.[1219] The Canadian Centre for Child Protection (CP3) also raised concern that the measure would only apply to services that have an existing method of age assurance.[1220]
- The Age Verification Providers Association (AVPA) and Yoti argued that this approach risked creating a disincentive for service providers to introduce age assurance technologies.[1221]

8.27 We respond to these specific concerns in the ‘How this measure works’ section (paragraphs 8.50 to 8.58).

### Age assurance methods

8.28 Yoti asked Ofcom to specify a range of methods that service providers may use to determine whether a user is a child or an adult (including self-declaration, profiling or marketing, or evidence of the age of users on similar sites).[1222] Similarly, WeProtect

[1213] As part of our November 2023 Consultation, we acknowledged that self-declaration was widely used by U2U services but could be easily evaded through deliberate false declaration We stated that where the only information service providers have is a user’s self-declared age, they should continue using it.

[1214] Barnardo’s response to November 2023 Consultation, pp.19-20; NSPCC response to November 2023 Consultation, p.35; Philippine Survivors Network response to November 2023 Illegal Harms Consultation, p.14; The British and Irish Law Education Technology Association (BILETA) response to November 2023 Consultation, pp.13-14.

[1215] UKSIC response to November 2023 Consultation, pp.17-18.

[1216] 5Rights Foundation response to November 2023 Consultation, p.26; Age Verification Providers Association (AVPA) response to November 2023 Illegal Harms Consultation, p.3; Canadian Centre for Child Protection (C3P) response to November 2023 Illegal Harms Consultation, p.23; Yoti response to November 2023 Illegal Harms Consultation, pp.18-19.

[1217] AVPA response to November 2023 Consultation, p.3; Yoti response to November 2023 Consultation, p.18.

[1218] Yoti response to November 2023 Consultation, p.18.

[1219] 5Rights Foundation response to November 2023 Consultation, p.26.

[1220] C3P response to November 2023 Consultation, p.23.

[1221] AVPA response to November 2023 Consultation, p.3; Yoti response to November 2023 Consultation, p.18.

[1222] Yoti response to November 2023 Consultation, p.3.

suggested it was "important" for privacy reasons, to provide users with a choice as to which age estimation tools they use to confirm their age online.[1223]

8.29 We discuss this in the 'How this measure works' section (paragraphs 8.50 to 8.58).

### Problems with "an existing means of identifying child users"

8.30 Two respondents asked for clarity on expectations around "an existing means of identifying" child users and suggested that it raised confusion on the extent to which service providers must confirm a child's identity.[1224] Yoti suggested that the term 'identifying' gives the impression that the full identity details of an individual are required. They argued that we should clarify that "only the data minimised age attribute is required from a hard identifier" (such as a formal identity document such as a passport) "or from a reusable digital ID app".[1225] The AVPA argued that the term 'identifying' suggests service providers must have full knowledge of the identity of a child, and suggested that the measure should instead specify whether a service provider "has an existing means of knowing which users are under 18 years old".[1226]

8.31 We respond to this feedback in the 'How this measure works' section (paragraphs 8.59 to 8.61).

## Feedback on how this measure works

8.32 UK Interactive Entertainment (Ukie) queried the applicability of network expansion prompts in the context of multiplayer gaming services. It suggested that gaming services often rely on 'matchmaking' players online, and raised concern that the measure could result in child users not being able to use functionalities that are crucial to multiplayer gameplay. For example, the removal of network expansion prompts may prevent child users from building teams in multiplayer games. Ukie also asked for clarification on the application of visibility of connection lists during gameplay, where players can see other players who are in the same gameplay as them.[1227] These concerns are addressed in the 'How this measure works' section (paragraphs 8.62 to 8.64).

## Feedback on the effectiveness of the measure

8.33 Several stakeholders also commented on the effectiveness of the safety defaults for child users.[1228]

- 5Rights Foundation, C3P and [✂] suggested that the effectiveness of this measure would materially increase if the settings in question were mandated (rather than being defaults) for some child users, particularly younger users.[1229]

[1223] WeProtect Global Alliance response to November 2023 Consultation, pp.19-20.

[1224] AVPA response to November 2023 Consultation, p.3; Yoti response to November 2023 Consultation, pp. 18-19.

[1225] Yoti response to November 2023 Consultation, p.18-19.

[1226] AVPA response to November 2023 Consultation, p.3.

[1227] UK Interactive Entertainment (Ukie) response to November 2023 Illegal Harms Consultation, p.25.

[1228] 5Rights Foundation response to November 2023 Consultation, pp.26-27; Barnardo's response to November 2023 Consultation, pp.19-21; BILETA response to November 2023 Consultation, p.14; C3P response to November 2023 Consultation, p.23-24; NSPCC response to November 2023 Consultation, p.35; Protection Group International response to November 2023 Consultation, p.10.

[1229] 5Rights Foundation response to November 2023 Consultation, pp.26-27; C3P response to November 2023 Consultation, p.23-24; [✂].

- Barnardo's did not suggest the settings should be permanent, but highlighted their research which found that children are pressured into being 'socially perfect'. They felt that children could 'deactivate' the default settings because of pressure from peers or perpetrators and remain at risk of abuse and exploitation.[1230] Similarly, older children aged 16+ in our engagement piece suggested that whilst it was important for children to have the option to disable settings, they felt that if they were younger, they would feel similar pressure to 'turn off' the defaults.[1231]
- The NSPCC broadly agreed with the settings as default, suggesting that it "rightly empowers children", but felt the measure placed a greater responsibility on children to manage their online experience.[1232]
- The British and Irish Law Education Technology Association (BILETA) said that the functionality settings should not be made permanent but did suggest a "protected time" period where a user could familiarise themselves with a service and the functionality before deciding whether to use it.[1233]
- Protection Group International felt that the default approach could be circumvented through children using search services to learn how to change the safety defaults.[1234]
- Children in our engagement piece commented on the effectiveness of the settings for the direct messaging functionality of services without a user connection functionality. They suggested that requiring children to actively confirm whether to receive a direct message from another user, could in some circumstances encourage curiosity, and may make them more likely to read a direct message (thereby questioning the extent of the effectiveness of the measure to make them consider not opening the direct message).[1235]

8.34 We respond to this feedback in the 'Effectiveness' section (paragraphs 8.77 to 8.86).

## Costs of the measure on users' online experiences

### Negative impact of the measure on child users' experiences

8.35 Snap raised concerns about the potential for the measure to have a negative impact on child users' online experiences. They felt it could either limit children's ability to connect and interact with other users or impact their overall wellbeing through a loss of connection:[1236]

- Snap suggested that network expansion prompts can support positive wellbeing by helping children make connections online, arguing that its 'quick add' functionality helps children feel connected. It highlighted its own research and research from Internet Matters that found that many children feel happy after spending time online. It also

[1230] Papamichail, M., Sharma, N., 2019. Left to their own devices: Young People, social media and mental health [accessed 06 November 2024]; Barnardo's response to November 2023 Consultation, pp.19-21.
[1231] Praesidio Safeguarding, 2024. Consulting children on proposed safety measures against online grooming [accessed 16 December 2024].
[1232] NSPCC response to November 2023 Consultation, p.35.
[1233] BILETA response to November 2023 Consultation, p.14.
[1234] Protection Group International response to November 2023 Consultation, p.10.
[1235] Praesidio Safeguarding, 2024. Consulting children on proposed safety measures against online grooming. [accessed 16 December 2024].
[1236] Snap response to November 2023 Illegal Harms Consultation, pp.17-20.

highlighted that being online enables children to keep in touch with friends while feeling that they are part of a group.[1237]

- Snap also suggested that network expansion prompts should be available for older children and where child users have multiple mutual contacts or contacts that are already in the child user's device address book.[1238]

8.36 While supportive of the measure, the NSPCC recognised that it could also have negative effects on children's online experiences, highlighting feedback from their 'young persons panel' that there are benefits to being able to see if someone who adds you has mutual connections.[1239] It suggested that the ability to view the connection lists of other child users may help them establish whether they know someone (through an assessment of mutual contacts) before deciding whether to accept a direct message from them. This was also echoed in our engagement with children.

8.37 We address these responses in the 'Costs and risks' section (paragraphs 8.98 to 8.104).

### Negative impact of the measure on adult users' experiences

8.38 Two respondents highlighted the negative impact of the measure on adult users' online experiences. An individual argued that it should not be impossible for people to contact children who they did not know, as such contact can lead to positive experiences, including making new friends.[1240] Another respondent also noted concerns about the measure's impact on adult users' freedom of expression and association.[1241] They felt adult users would be less willing to engage with children online for legitimate reasons because of the measure.[1242]

8.39 We respond to these issues in 'Costs and risks' and 'Rights impact' sections (paragraphs 8.105, 8.106, 8.110 and 8.111).

## Feedback on who the measure applies to

### Broadening the scope of the measure

8.40 In response to the November 2023 Consultation, several stakeholders suggested that the measure's scope should be broadened to include more services.[1243] Some civil society organisations suggested extending the measure to all services at medium or high risk of grooming irrespective of size, or applying it to all services at medium or high risk of any kinds of illegal harms.[1244] One ID Ltd felt that the measure would also risk "the majority of platforms" being out of scope.[1245]

[1237] Snap, 2024. New Research from University of Chicago's NORC Shows Communicating Online with Friends Brings Happiness for Teens & Young Adults. [accessed 07 November 2024]; Internet Matters, 2024. Children's Wellbeing in a Digital World. Our digital wellbeing research for 2024. [accessed 07 November 2024].

[1238] Snap response to November 2023 Consultation, p.19.

[1239] NSPCC response to November 2023 Consultation, p.36.

[1240] Julia H response to November 2023 Illegal Harms Consultation, p.3.

[1241] Name Withheld 3 response to November 2023 Illegal Harms Consultation, p.17.

[1242] Name Withheld 3 response to November 2023 Consultation, p.17.

[1243] 5Rights Foundation response to November 2023 Consultation, p.26; C3P response to November 2023 Consultation, p.23; Molly Rose Foundation response to November 2023 Illegal Harms Consultation, p.34; One ID Ltd response to November 2023 Consultation, p.2.

[1244] 5Rights Foundation response to November 2023 Consultation, p.26; C3P response to November 2023 Consultation, p.23; Molly Rose Foundation response to November 2023 Consultation, p.34.

[1245] OneID Ltd response to November 2023 Consultation, p.2.

8.41 However, Mid Size Platform Group (MSPG) agreed with our proportionality assessment and suggested this measure (and the further measure in this chapter) should be compatible with a variety of services and business models.[1246]

8.42 Although agreeing with the services we proposed applying this measure to, the Children's Commissioner for England argued that the safety defaults would be more effective if they were implemented as standard across "all services" and "all accounts".[1247]

8.43 We respond to this feedback in the 'Who this measure applies to' section (paragraphs 8.119 to 8.136)

### Displacement of users onto other services

8.44 Barnardo's and Protection Group International raised concerns about who this measure applies to, arguing that it could displace perpetrators and child users (particularly those who are younger) onto other services that would originally be considered less risky. They felt that such services (including those that are smaller and less 'risky') could fall outside the scope of this measure.[1248]

8.45 We address this theme in the 'Who this measure applies to' section (paragraphs 8.131 to 8.132).

## Our decision

8.46 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation. We have made minor clarificatory changes in response to the feedback outlined in the previous section:

- Our measure sets out that providers should apply the relevant safety defaults to child user accounts. In our November 2023 Consultation, we proposed the measure would apply to providers with "existing means of identifying child users of the service concerned".[1249] We have updated this in our final Codes to refer to providers with an "existing means to determine a user's age or age range of a particular user of the service concerned" and have added a definition within the codes to explain our expectations of what such means may include. We discuss this change further in the 'How this measure works' section.
- We have made some minor changes to the wording and relevant definitions for our measure, to clarify our expectations for the implementation of this measure. These changes include definitions and descriptions for network expansion prompt functionality, connections, direct messaging functionality and location information.

8.47 The full wording of the measure can be found in our Illegal Content Codes of Practice for U2U services, in which we refer to this measure as ICU F1. As we explain below, we are considering consulting in the future on proposals which would require services which pose a high risk of grooming to use Highly Effective Age Assurance to determine who is a child for the purposes of implementing this measure. However, in the interim we will apply the measure to services with an existing means to determine a user's age or age range.

[1246] Mid Size Platform Group (MSPG) response to November 2023 Illegal Harms Consultation, p.10.
[1247] Children's Commissioner for England response to November 2023 Consultation, pp.22-23.
[1248] Barnardo's response to November 2023 Consultation, p.20; Protection Group International response to November 2023 Consultation, p.10.
[1249] See Annex 7 (A7.1) of the November 2023 Illegal Harms Consultation.

## Our reasoning

### How this measure works

8.48 We recommend that all providers of U2U services with a high risk of grooming and all providers of large U2U services with a medium risk of grooming should implement the following safety defaults (where they have an existing means of determining a user's age or age range and where relevant functionalities exist).

- Child users should not be presented with prompts to expand their network of friends or be included in network expansion prompts presented to other users.
- Child users should not be included in connection lists. Child users' connections lists should not be displayed to other users.
- Where services have a user connection functionality which allows users to connect with one another (for example, to become 'friends'), users should not be able to send direct messages to child users without first establishing a specified connection.[1250]
- Where services have no user connection functionality, child users should be provided with a means of actively confirming whether to receive a direct message from a user before it is visible to them. However, if direct messaging is a necessary and time-critical element of another functionality, then in addition to actively confirming, child users should be informed about receiving direct messages during the use of the other functionality before the child user commences interaction associated with that other functionality.
- Automated location information displays, which automatically create and display location information for child user accounts, should not display child users' location information.

8.49 We provide further details about these features and functionalities in the 'Benefits' section (paragraph 8.73).

#### Determining a user's age or age range

8.50 As the recommendation only covers applying the safety defaults to child user accounts, it follows that the more effectively service providers can determine which of their users are children, the greater the efficacy of the measure in mitigating the risk of grooming harm. Services currently use a variety of methods to determine whether users are adults or children (subject to applicable data protection and privacy laws). These include (but are not limited to) age verification, age estimation, and self-declaration processes.[1251]

8.51 As we set out in the November 2023 Consultation – and as a number of respondents noted – a significant number of children lie about their age when using self-declaration to create

[1250] A specified connection is the type of connection which must be in place with a child user in order for a direct message to be received by that child user. For more details on what constitutes a specified connection see the Illegal Content Codes of Practice for U2U services, ICU F1.5

[1251] We recognise Yoti's concern regarding Ofcom specifying other ways in which service providers may determine whether a user is a child. Service providers may use various methods to establish the age of a user and our list is not exhaustive but include those that we consider are commonly used among providers. Furthermore, Section 230 (1)-(4) of the Act specifies definitions for age verification, age estimation, and self-declaration.

social media accounts.[1252] A respondent also noted a concern that the measure could be easily circumvented through self-declaration in the absence of an age assurance requirement.[1253]

8.52 Provisionally, we considered that the introduction of age assurance would strengthen the effectiveness of the safety defaults measure if introduced by service providers, as it would provide increased accuracy in determining a user's age or age range.

8.53 Our research supports this, finding that a third of child respondents aged eight to 17 who had a social media profile pretended to be aged 18 or over.[1254] This suggests that our measure would not be implemented for all children's accounts where a service relies on self-declaration to determine a user's age. Self-declaration is therefore an imperfect method to determine the age of a user and is not capable of being 'highly effective'.[1255]

8.54 Following our November 2023 Consultation, Ofcom included proposals on highly effective age assurance ('HEAA') in the December 2023 Consultation on our guidance for service providers publishing pornographic content on their online services ('Part 5 guidance'). Age assurance proposals were also included for U2U services in our May 2024 Consultation. However, our expectations around HEAA will not be finalised at the time we publish these Illegal Content Codes of Practice for U2U services.[1256]

8.55 Therefore, rather than delay the introduction of the safety defaults measure, we proposed in our November 2023 Consultation that we should initially introduce the measure with a stipulation that services should only be in scope if they have an existing means of identifying child users, whether that is a form of age assurance or another method.

**How this measure will interact with the Children's Safety Codes**

8.56 From the point at which the Children's Safety Codes come into effect, relevant providers of U2U services will need to implement HEAA (where applicable) to comply with the child safety duties. From this point, service providers that are in scope of both the HEAA measures in the Children's Safety Codes and our safety default measure will need to use HEAA to determine who is a child for the purposes of implementing the measure.

8.57 Prior to the children's safety duties coming into force however, there will be a short period where service providers will have discretion as to what method they use to determine a user's age or age range for the purposes of implementing measures in this chapter.[1257] We

[1252] Barnardo's response to November 2023 Consultation, p.19-20; BILETA response to November 2023 Consultation, p. 13-14; NSPCC response to November 2023 Consultation, p. 35; Philippine Survivors Network response to November 2023 Consultation, p. 14.

[1253] UKSIC response to November 2023 Consultation, pp.17-18.

[1254] Ofcom, 2022. A third of children have false social media age of 18+. [accessed 07 November 2024].

[1255] See Child Safety Duties, Section 12(6) of the Act; Part 5: Section 81(3) of the Act; and Codes of Practice Principles, Schedule 4, paragraph 12 of the Act.

[1256] The Children's Safety Codes and the Part 5 guidance form phase 2 of Ofcom's implementation of the Online Safety Act 2023, whereas the measures in this chapter fall under phase 1. In January 2025, we intend to issue our final Part 5 guidance and we also intend to publish our final Children's Safety Codes in April 2025. We expect that the child safety duties will come into effect from July 2025.

[1257] From the point at which the children's safety duties come into effect (which we expect around July 2025), relevant providers of U2U services will need to implement HEAA (where applicable) to comply with the children safety duties. Where relevant providers are using HEAA to comply with children safety duties, we expect that they will use HEAA to apply the measures in this chapter. Prior to this however, there will be a short period where service providers will have discretion as to what method they use to determine a user's age or age range for the purposes of implementing measures ICU F1 and F2.

consider this option preferable to delaying the introduction of the measures as they offer significant benefits to children from grooming risks, as we explain in the 'Benefits' section for this measure.

8.58 We further recognise that it is possible that some service providers that are in scope of the measures in this chapter may not be in scope of the children's safety duties. As things stand, these service providers would not be expected to use HEAA once the Children's Safety Codes come into force. We appreciate that such a situation could limit the effectiveness of the measures recommended in this chapter. We will closely monitor how many services may be out of scope of the HEAA requirement and review any future evidence surrounding this issue in due course. Should we conclude that a material number of services in scope of these measures are not captured by the HEAA requirements in the Children's Safety Codes, we will consult in the future on expanding our HEAA measures to close this gap.

### "Identifying" child users

8.59 On a related theme, some stakeholders asked for clarity surrounding our expectations of "an existing means of identifying" child users, suggesting there was confusion on the extent to which service providers must confirm a child's identity, rather than a child's age.[1258]

8.60 We acknowledge respondents' concerns that the term may be confusing for service providers. We also note that the term 'identifying' does not accurately describe the available methods associated with age verification and estimation. Many of these methods assess if a user is an adult, and as a result of doing so, services employing these methods infer that any user not determined to be an adult is a child.

8.61 In light of this feedback, we have updated the Codes to state that services should apply the measure if they have an "existing means of determining the age or age range of a particular user of the service concerned".[1259] We consider this better describes the technologies used to confirm whether a user is a child or an adult, as well as clarifying the confusion around the term 'identifying'.

### Applicability of the network expansion prompt settings

8.62 As described in paragraph 8.32, UKIE queried the applicability of network expansion prompts in the context of multiplayer gaming services. It also raised a concern that the measure could result in child users not being able to use functionalities that are crucial to gameplay.[1260]

8.63 Our measure is intended to restrict the recommendations of 'connections,' typically associated with online interactions like 'following' or 'subscribing' on all types of services, including gaming services.[1261] However, this does not include gameplay set-up situations, where players are recommended to other players to temporarily come together to play a game. We expect providers of gaming services to consider the characteristics of their gameplay and whether it establishes a connection between users when implementing the measure.

[1258] AVPA response to November 2023 Consultation, p.3; Yoti response to November 2023 Consultation, pp.18-19.

[1259] See Illegal Content Codes of Practice for U2U services, ICU F2.1.

[1260] Ukie response to November 2023 Consultation, p.25.

[1261] See definition for 'connection' in Illegal Content Codes of Practice for U2U services.

8.64 Although this measure does not apply to certain functionalities and features related to game play, our evidence suggests that children face a significant risk of harm from grooming on gaming services. As such, we expect providers to consider the applicability of the measure on their service for other relevant functionalities.[1262]

## Benefits

### Prevalence and severity of online CSEA

8.65 While it is not possible to accurately determine the full scale of online grooming in the UK, the evidence demonstrates that it is both a widespread and growing issue, causing harm to a substantial number of children in the UK every year.[1263] Furthermore, estimates are likely to be a significant underestimate of the true extent of harm.

8.66 The NSPCC reports there were nearly 34,000 recorded online grooming crimes against children in the six years up to 2023, and an 82% increase in sexual communication with a child offences between 2017 to 2018 and 2022 to 2023 – but these prosecutions will account for only a small proportion of actual and attempted grooming activity.[1264] Retrospective studies with young adults reflecting on their experiences as children indicate that the proportion of children who have online experiences indicative of potential grooming can be as high as one in four.[1265]

8.67 The prevalence of self-generated intimate imagery (SGII) can also be indicative of the potential scale of grooming because SGII can involve children being coerced into sending intimate images. In 2023, SGII featured on 92% of sites against which the Internet Watch Foundation (IWF) took action.[1266] There is also a growing trend of SGII being used as a method to blackmail children for financial payments.[1267]

8.68 Online CSEA causes severe and often lifelong harm to victims, and its effects extend to other children, communities, wider society, and public services.[1268] This includes long-term mental health challenges, such as suicidal ideation, depression, anxiety, and post-traumatic stress symptoms.[1269] The use of deception and coercion by perpetrators can lead victims

[1262] See Register of Risks chapter titled 'CSEA' for more details.

[1263] For a detailed exploration of these issues and the harm caused to children, see the Register of Risks chapter titled 'CSEA'.

[1264] NSPCC, 2024. 82% rise in online grooming crimes against children in the last 5 years. [accessed 07 November 2024].

[1265] Greene-Colozzi, E., Winters, G., Blasko, B., and Jeglic, E., 2020. Experiences and Perceptions of Online Sexual Solicitation and Grooming of Minors. A Retrospective Report. *Journal of Sexual Abuse*, 29:7, p.836-854. [accessed 07 November 2024] The study was of 1,133 undergraduate college students at two public institutions in the United States, who were asked about their experiences when under 18.

[1266] The IWF found that of the 275,652 webpages it acted on during 2023, 92% were assessed as containing SGII. Source: IWF, 2023 'Self-generated' child sexual abuse [accessed 07 November 2024].

[1267] For a detailed exploration of this emerging area of harm, see the Register of Risks chapter titled 'CSEA'.

[1268] Owens, J. N., Eakin, J. D., Hoffer, T., Muirhead, Y., and Shelton, J. L. E. 2016. Investigative aspects of crossover offending from a sample of FBI online child sexual exploitation cases. *Aggression and Violent Behavior*, 30, 3–14 [accessed 07 November 2024]; C3P, 2017. Survivors' survey: Executive summary 2017, pp. 28-29. [accessed 07 November 2024]; The Independent Inquiry into Child Sexual Abuse (IICSA), 2022. I will be heard: Victims and survivors' experiences of child sexual abuse in institutional contexts in England and Wales, pp. 104-112. [accessed 07 November 2024]; IICSA, 2022. Part G: "The impact of child sexual abuse" in The Report of the Independent Inquiry into Child Sexual Abuse, [accessed 07 November 2024];

[1269] Joleby, M., Lunde, C., Landstrom, S., and Jonsson, L. S. 2020. "All of me is completely different": Experiences and consequences among victims of technology-assisted child sexual abuse. *Frontiers in Psychology*, 11, 606218. [accessed 07 November 2024].

and survivors to experience self-blame and a loss of trust, which can have a prolonged impact on their relationships with their community and their interpersonal relationships.[1270]

8.69 There is also a risk that once a child has been groomed, the abuse can extend to other children, including siblings and friends.[1271] This implies that reducing one instance of grooming could subsequently reduce harm to more than one child.

8.70 Ultimately, the prevalence and severity of online CSEA, including the grooming of children by adults, is a serious harm and a priority to address. We have taken this into account in the development of the final measures and we describe the evidence more fully in the Register.[1272]

### Benefits of our measure

8.71 Based on the evidence outlined, we consider that the measure proposed will materially disrupt the grooming process that can take place on U2U services. This will occur principally by reducing the current ease with which perpetrators can communicate with or identify children.

8.72 We are aware that many child users enjoy using online services to connect with their friends and likeminded communities. However, evidence also demonstrates that perpetrators deploy a range of techniques to contact children and will often exploit certain functionalities on U2U services to identify and target children.

8.73 In the following paragraphs, we discuss how these functionalities may be used to facilitate online grooming by considering each group in turn. We then consider the measure's effectiveness in disrupting this behaviour and the benefits that it will bring to children.

- **Restriction of network expansion prompt functionalities**. Network expansion prompt functionalities use a recommender system to suggest other users to connect with, based on the service's knowledge of its users. This can include specific users who have similar interests, who are close geographically, who attend the same school or workplace, or with whom the user has a mutual connection. As identified in our Register, these functionalities can play a key role in facilitating grooming, and perpetrators often utilise network expansion prompts to identify multiple children within a similar network.[1273] Removing children from network expansion prompts will reduce the speed and ease with which potential perpetrators can target and contact children using a service.[1274] Furthermore, this will make it harder for perpetrators to find and connect with children they do not know using either a targeted or 'scattergun' approach.[1275] We consider that these restrictions could reduce the amount of grooming initiated online and that this

[1270] Schmidt, F., Bucci, S., and Varese, F. 2023. Understanding the prolonged impact of online sexual abuse occurring in childhood. *Frontiers in Psychology*, [accessed 08 October 2024].

[1271] IICSA, 2020. Part D.3: "Victims and survivors" in The Internet Investigation Report March 2020. [accessed 02 October 2024].

[1272] We give a more detailed overview of the evidence on grooming and its impact in the Register of Risks chapter titled 'CSEA'.

[1273] See Register of Risks chapter titled 'CSEA'.

[1274] As we explained in our November 2023 Consultation, given that perpetrators may use mutual connections to generate trust, not disclosing mutual connections would also reduce the risk of children connecting with perpetrators who may have connections in common.

[1275] The 'scattergun approach' is a process in which perpetrators attempt to connect with child users by contacting multiple children in a short period of time.

could in turn lead to a reduction in the amount of sexual abuse occurring as a result (relative to if this measure were not in place).

- **Restriction of connection lists functionalities**. As explained in our November 2023 Consultation, on some services, a user's connections are visible to other users via their profile. This includes features such as 'friends', 'followers', 'subscribers', or indications of mutual connections. As identified in our Register, such functionalities are exploited by those seeking to groom children for the purposes of sexual abuse.[1276] There are a number of ways in which this happens. We understand that once perpetrators have connected with a child, they sometimes use their 'friend' lists to identify further children to target. Related to this, we understand that perpetrators sometimes use mutual connections to increase children's confidence in communicating with them by facilitating a false sense of trust – in other words, once a child sees that one of their friends is connected to a perpetrator, they are more likely to trust that perpetrator. Furthermore, evidence suggests that connection lists can be used by perpetrators in coercion and blackmail cases involving children. We understand that blackmail in particular is commonly used to generate CSAM imagery. This is facilitated if the child knows that the perpetrator has knowledge of and the ability to communicate with the child's family and friendship groups.[1277] Ensuring child users' connection lists are not visible to other users (and that child users are not included in the connection lists of other users) will make it harder for perpetrators to exploit connection lists in the ways described above.

- **Restriction of direct messaging functionalities**. Direct messaging functionalities allow message exchanges between two users in an interface that cannot be viewed by other users.[1278] As outlined in the Register, these functionalities are also exploited for grooming offences, as perpetrators often develop relationships with children away from public view and parental supervision.[1279] Evidence suggests that in nearly three quarters of cases (74%) where children are contacted online by someone they do not know in person, this contact involves private messaging.[1280] Perpetrators often use direct messaging functionalities to send children unsolicited messages, either for the purposes of committing a grooming offence or to engage in other forms of communication that could increase the risk of harm to a child in relation to other offences.[1281] Restricting direct messaging functionalities will make it much harder for perpetrators to do this, thereby materially reducing the amount of grooming that occurs.

- **Restriction of location information**. Location information may be displayed or shared on U2U services either automatically by the provider of the service (through particular functionalities such as 'live location' functionalities), through the automated display of location in user profiles or shared content, or through manual input by users on shared

[1276] See Register of Risks chapter titled 'CSEA'.

[1277] For further detail see the Register of Risks chapter titled 'CSEA', with specific focus on the 'User connections' and 'Network recommender systems' subsections for grooming.

[1278] Direct messaging is a functionality allowing a user to send and receive a message to one recipient at a time and which can only be immediately viewed by that specific recipient.

[1279] See Register of Risks chapter titled 'CSEA'.

[1280]Office for National Statistics, 2021. Children's online behaviour in England and Wales: year ending 2020. [accessed 15 October 2024].

[1281] This includes (but is not necessarily limited to) harassment, stalking, threats and abuse, hate, controlling or coercive behaviour (CCB), or even terrorism. We set out further details how our measures can reduce other risks of harm in paragraphs 8.74 to 8.76.

content. As outlined in the Register, the sharing of location information may provide perpetrators with the necessary knowledge of places frequently visited by a child user, such as their home, school, or other locations.[1282] This knowledge may enable offline contact, which could lead to contact sexual abuse.[1283] Automated sharing is of particular concern because children may be unaware that their location is being shared. Restricting the automatic display of location information related to child user accounts may reduce the risk of children being live-tracked without their knowledge and a subsequent risk of contact sexual abuse. We address this issue further in our Register.

### The measure's ability to reduce other risks of harm

8.74 We also consider that this measure will assist providers in mitigating risks of harm related to other offences, which we see as an additional benefit to the measure.

8.75 The measure will impact the ability of perpetrators to use a broad range of functionalities to identify, connect and communicate with child users, such as direct messaging, user connections, user profiles, and sharing location information. The evidence in our Register shows that children are at particular risk of various kinds of illegal harm where these functionalities play an important role. This includes harassment, stalking, threats and abuse, hate, controlling or coercive behaviour (CCB), and terrorism.[1284]

8.76 The measure will make it more difficult for perpetrators to use these functionalities to target child users for the purpose of committing such offences. This may assist providers in complying with their duties to mitigate the risks of harm related to these offences, as well as CSEA offences. Therefore, we recommend that the safety defaults measure is included as part of the Codes for CSEA and other duties given that service providers who implement measures to address a risk of grooming on their service are likely to simultaneously address the risk factors related to other harms.[1285]

## Effectiveness

8.77 Some respondents queried the extent of the measure's effectiveness. One respondent suggested that it placed too great a responsibility on children to keep themselves safe.[1286] Another argued that children could be pressured into 'turning off' the default settings, thereby remaining at risk of grooming.[1287] Other respondents also suggested that the

[1282] Please see the Register of Risks chapter titled 'CSEA'.

[1283] Please see the Register of Risks chapter titled 'CSEA'.

[1284] See the Register of Risks chapters for the kinds of illegal harm identified for further discussion of how children specifically can be at increased risk of harm in relation to these offences. For example, regarding CCB, we know these functionalities are used to identify and target children to gain access to victims, and target teenage girls who may be the direct victims of CCB themselves.

[1285] We note the Independent Reviewer of Terrorism Legislation response to the November 2023 Consultation, p.7, and a suggestion that the measures recommended in this chapter are also extended to the Terrorism Code. We are not extending the measure to the Terrorism Code, as there is limited evidence on child terrorism activity in the UK. However, we note that there are indicators that children may be at increased risk of radicalisation due to the time they spend online and consider that the measure may mitigate the radicalisation of children in some cases where the targeted functionalities are used in a similar way to commit grooming. We set out more details regarding evidence which suggests that children may be at increased risk of radicalisation due to the amount of time they spend online in the Register of Risks chapter titled 'Terrorism'.

[1286] NSPCC response to November 2023 Consultation, p.35

[1287] Barnardo's response to November 2023 Consultation, p.20.

effectiveness of the measure would increase if the features were mandated for some child users, particularly younger children.[1288] We consider these points below.

### Changing the defaults

8.78 As explained in our November 2023 Consultation, we recognise that the effectiveness (and therefore the benefits) of our measures would be reduced if a material proportion of children change the default settings. Some may do so voluntarily, particularly older children who may be better informed of the safety risks associated with disabling the defaults. In other cases, children may be pressured by peers or perpetrators to change the defaults.

8.79 Even though some children may change the default, the evidence suggests that the majority will not. Research indicates that when presented with pre-set courses of action, people are generally more likely to stick with the default option than choose another one.[1289] Our own behavioural research into user content controls also found default settings to be 'sticky' in that, even when prompted, few participants moved away from the content control default setting.[1290] There is also evidence across a range of different contexts that suggests that setting defaults is effective at influencing choices and behaviours.[1291]

8.80 Our engagement with children found them to be broadly supportive of the default approach. When asked if they would 'turn off' the default settings, most children indicated that they would not and welcomed the ability to decide.[1292] We consider that a default approach is therefore 'choice preserving', meaning children can decide if they wish to change their settings – to turn them off, but also to turn them back on again if they so wish. We also consider the risks of children switching off the settings would be somewhat mitigated through our other measure, recommending that child user accounts are provided with support at the point of doing so (as we later discuss).

### Permanently disabling features

8.81 We recognise stakeholder feedback that permanently disabled features for younger children could be beneficial. This aligns with feedback we received in response to the May 2024 Consultation calling for age-specific measures.[1293] Ofcom is in the process of analysing these responses and determining its approach to age-specific measures and will confirm its final approach in the Protection of Children Statement. Following this Statement, and subject to evidence on age assurance methods and age-related risk, we will consider

[1288] 5Rights Foundation response to November 2023 Consultation, pp.26-27; C3P response to November 2023 Consultation, pp.40-42; [✂].

[1289] Thaler, R. H., Sunstein, C. R., and Balz, J. P., 2013, Choice architecture. In E. Shafir (Ed.), *The behavioral foundations of public policy* (pp. 428-439). Princeton, NJ: Princeton University Press.

[1290] This study involved adults rather than children, but we are not aware of any evidence which suggests that children would be likely to respond in a very different way as adult participants. Ofcom, 2024. 'Behavioural insights to empower social media users. Testing tools to help users control what they see', Behavioural Insights Discussion Paper. [accessed 25 November 2024].

[1291] Jachimowicz, J., Duncan, S., Weber, E., and Johnson, E., 2019. When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, 3(2), pp.159-186. [accessed 26 November 2024]; Mertens, S., Herberz, M., Hahnel, U.J.J., and Brosch, T. (2021). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioural domains. *Proceedings of the National Academy of Sciences*, 119(1). [accessed 30 October 2024].

[1292] Praesidio Safeguarding, 2024. Consulting children on proposed safety measures against online grooming.[accessed 16 December 2024].

[1293] 5Rights Foundation response to May 2024 Consultation, pp.11-14; NSPCC response to May 2024 Consultation, pp. 28-33.

permanently disabling features for younger users in reference to Illegal Content Codes measures.

8.82 Provisionally, we consider that a measure recommending permanently disabled functionalities would also have a significant impact on children's rights to freedom of expression and association. There may be legitimate reasons why older teenagers in particular wish to turn off the default settings. There is an argument that they should have the ability to make an informed decision to do so, even if this places a responsibility on the child user to manage their online experience.[1294]

8.83 Additionally, permanently disabling any of these functionalities would also restrict a child user's ability to develop an enterprise which relies on monetising content. Based on our current evidence on default settings, we consider this level of interference with children's rights and online experiences to be disproportionate, but we will continue to keep this under review as part of our monitoring of the measure's effectiveness and may return to evaluate it at a later date.

### Other elements impacting the effectiveness of the measure

8.84 We also appreciate BILETA's suggestion that there could be a "protected time period" for children before they gain full use of features and functionalities on a service (such as restricted access to direct messages).[1295] If we were to introduce any new requirements, it would change our impact assessment as discussed above, and we do not have the evidence at this time to review the measure in such a manner. We will continue to follow developments in these areas and any evidence that may inform our future work.

8.85 We also note children's comments surrounding the effectiveness of the direct messaging setting for services without a user connection functionality.[1296] While we appreciate children may be "curious" to read a message as result of the setting, the alternative of having no setting in place, would see children receiving messages with no barrier. Therefore, if the guardrails we recommend help some children consider this message before deciding whether to read it, we consider that a success. Additionally, if a child decides to disable the related setting, they will receive a supportive message reminding the child that this is the first communication with that user (which could cause a reassessment of that choice).

8.86 In summary, our evidence demonstrates the significant harm that arises from the grooming of child users. Our analysis shows that this measure should disrupt the grooming process and therefore has the potential to reduce this harm and could help reduce the sexual abuse of children as a result. This assessment is supported by various civil society respondents, the Children's Commissioner for England, the ICO, and the children who were surveyed as part of their engagement with the measures.[1297] Given the severe impact of grooming and

[1294] NSPCC response to November 2023 Consultation, p.35.

[1295] BILETA response to November 2023 Consultation, p.14.

[1296] Praesidio Safeguarding, 2024. Consulting children on proposed safety measures against online grooming. [accessed 16 December 2024].

[1297] 5Rights Foundation response to November 2023 Consultation, p.26-27; Children's Commissioner response to November 2023 Consultation, p. 22; ICO response to November 2023 Consultation, p.20; NSPCC response to November 2023 Consultation, p.34-37; UKSIC's response to November 2023 Consultation, p.17; We Protect Global Alliance response to November 2023 Consultation, p.19-20; Praesidio Safeguarding, 2024. Consulting children on proposed safety measures against online grooming. [accessed 16 December 2024].

how prevalent it is, we conclude that the benefits associated with this measure will be very significant.

## Costs and risks

8.87 In our November 2023 Consultation, we explained that this measure would give rise to:

- the direct costs of modifying services driven by engineering costs, overheads, and coordination costs;
- the indirect costs to service providers resulting from lost revenue;
- the indirect costs to child users of a service as a result of lost functionalities; and
- the indirect costs to adults from making it more difficult to contact children (when this is done with good intentions and benefits both the adult and the child)

8.88 After considering stakeholder responses, our assessment of these costs remains largely unchanged. In the following paragraphs, we describe our approach to costs and the relevant stakeholder responses.

### Direct costs to service providers

8.89 We consider the direct costs of modifying a service to be driven by (i) the engineering costs needed to set up the measure and (ii) the overhead and coordination costs needed to review the impacts on the service.

8.90 We had limited feedback in our November 2023 Consultation on the specific direct costs of implementing the measure and have kept the cost assumptions and estimates largely unchanged.[1298]

8.91 We estimate the one-off upfront engineering cost associated with the implementation of all recommended default settings under the measure to be approximately £10,000 to £115,000, made up of two to 12 months' worth of staff resources (split equally between software engineering staff and other professional occupation staff such as project managers).[1299]

8.92 We estimate the one-off upfront overhead and coordination costs associated with any change to the frontend or backend systems of a service to be £0 to approximately £210,000, made up of zero to 24 months' worth of professional occupation staff resources.[1300]

8.93 We estimate that the total one-off upfront cost (including both engineering, overheads, and coordination costs) will range from £10,000 to £325,000.

8.94 The variation in costs is driven by differences in both the size of a service and of the provider's existing systems. Costs will likely be towards the lower end of the estimated range if a service provider already gives users options to limit their appearance in network expansion prompts and to control the visibility of their connections. Costs may be higher for

[1298] We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS. However, since our cost estimates are rounded, the estimates may not necessarily have changed when using the updated wage assumptions. We received some general feedback on the cost assumptions (such as salary assumptions) that are fed into these costs. We consider that feedback and our updated wage assumptions in Annex 5.

[1299] Based on our standard assumptions for labour costs set out in Annex 5.

[1300] Based on our standard assumptions for labour costs set out in Annex 5.

service providers developing the recommended measure from scratch. This may involve upgrading the backend of their website (including databases and data storage) as well as upgrading website user interfaces. Similarly, the costs associated with these modifications will likely be lower for service providers using an off-the-shelf tool like WordPress to build and maintain their services. Costs may also be higher for providers who need to modify underlying code and service infrastructure. We also expect overhead and coordination costs to be correlated with service provider size because large providers employing thousands of staff will require significant review, communication, and legal processes to implement the modifications recommended under this measure.

8.95 In addition to the upfront costs, we assume the ongoing costs of reviewing and monitoring the measure to be 25% of the upfront costs on an annual basis, ranging from £2,500 to £81,250.[1301]

### Indirect costs to service providers

8.96 We acknowledge that service providers may incur indirect costs resulting from lower user activity and subsequent lost revenue. We distinguish between (i) reductions in revenue resulting from a reduction in the occurrence of illegal activity on a service and (ii) reductions in revenue resulting from a reduction in legitimate activity on a service. We do not consider (i) as a relevant cost to factor into our impact assessment. We consider that (ii) may arise, for example, if users spend less time on a service leading to potential loss in advertising revenue for the provider. We consider that this will be manageable for service providers who already implement functionalities similar to those recommended under this measure.[1302] However, this may have a more substantial effect on smaller or newer service providers (who may find that their online growth depends on the presence of user network effects).

8.97 We also consider the recommendations outlined under this measure to have potential countervailing positive effects on user engagement on services. Reductions in grooming attempts and other forms of unsolicited contact from strangers (including harmful contact such as receiving unsolicited sexual images) may result in child users feeling more comfortable using a service. This may lead to an increase in user activity and subsequently additional revenue on the service.

### Indirect costs to child users of a service

8.98 We acknowledge that the measure may also have indirect costs to child users. It may have an adverse impact on how children interact online, including affecting their ability to make new friends online. Less frequently, it could also affect the ability of children to monetise their online presence by attracting users to their content.

[1301] Based on our standard assumptions for ongoing maintenance of software changes set out in Annex 5.

[1302] For example, on TikTok, the ‘suggest your account to others’ feature is turned off by default for users aged under 16 and needs to be actively enabled in privacy settings. Source: TikTok, 2023, New features for teens and families on TikTok. [accessed 10 October 2024]. On Instagram, teen accounts can be set to private, and adults exhibiting ‘potentially suspicious behaviour’ are restricted from seeing teen accounts in ‘Suggested Users’ or discovering teen content in ‘Reels’ or ‘Explore’. Source: Instagram, 2021, Continuing to Make Instagram Safer for the Youngest Members of Our Community. [accessed 15 October 2024]. Snapchat limits discoverability of teen accounts on their platform to people users are “likely [to] know” such as where there is a mutual connection. In addition, the friend lists of under 18 accounts are always private on Snapchat. Source: Snap, 2022, Parent’s Guide: Snapchat’s Family Center. [accessed 15 October 2024].

8.99 Some stakeholders noted concerns on the potential negative effect of the measure on child users' experience, specifically through the removal of network expansion prompts and the removal of connection lists of child users.[1303]

8.100 Our measure does not prevent children from searching for (and connecting with) other users online. Children will still be able to search for the names or usernames of family or friends to connect with others. They may also be able to use other connection request processes (such as scanning QR codes or user ID search) to connect with new friends online.

8.101 We note NSPCC's feedback that children may benefit from seeing the connection lists of other child users as it may help them establish whether they know someone (through an assessment of mutual contacts).[1304] However, our evidence suggests that where a perpetrator has mutual connections with a child user, it may appear to the child that the perpetrator is known to their social network. This can create a false sense of 'relationship' and 'trust' towards the perpetrator.[1305]

8.102 We also note Snap's suggestion that network expansion prompts should be available for older children and based on contacts in a user's device.[1306] As discussed in 'age of children covered by the measures', our evidence suggests that a significant risk of harm from grooming is present for children of all age range, including older children. We have therefore recommended that our measures and the relevant functionalities apply to all children under 18 to ensure that older children also have protections against CSEA.

8.103 Furthermore, we do not consider that it is appropriate to allow an exception for network expansion prompts based on contacts in a user's device. We consider there may still be risks associated with expansion prompts using existing contacts because perpetrators use multiple services in the grooming process as well as establishing the first contact in an offline space before moving the child to online spaces.[1307]

8.104 Importantly, the measure allows for children to have control over the settings as discussed in 'Effectiveness'. This allows them to make ongoing choices about their own online experiences. To some extent, children can mitigate any indirect costs they experience because they can change the default settings if they wish (although they may still be affected by other child users' choices). We recommend that children who change the settings should be presented with supportive information to help them make informed choices (see our measure on support for child users).

### Indirect costs to adults

8.105 We acknowledge that the measure may have indirect costs for adult users of a service by making it harder for them to connect with children online. However, as discussed in paragraphs 8.110 and 8.111, this measure will not prevent adult users who have a legitimate reason to connect with a child user from doing so. We consider that network

[1303] NSPCC response to November 2023 Consultation, p.36; Snap response to November 2023 Consultation, pp.17-20.

[1304] NSPCC response to November 2023 Consultation, p.36.

[1305] See Register of Risks chapter titled 'CSEA'.

[1306] Snap response to November 2023 Consultation, p.19.

[1307] See Register of Risks chapter titled 'CSEA'; Joleby, M., Lunde, C., Landström, S., and Jonsson, L. S. (2020). "All of me is completely different": Experiences and consequences among victims of technology-assisted child sexual abuse. Frontiers in psychology, 11, 606218.

expansion prompts and connection lists are not the only way for users to identify potential connections and that adults will still be able to connect with children they know.

8.106 While we recognise this measure could make it harder for adults to connect with children online, we expect it to bring significant benefits in terms of reducing the risk of grooming by adults seeking to connect with children that they do not know online.

## Rights impact

### Freedom of expression and freedom of association

8.107 As explained in 'Introduction, our duties, and navigating the Statement', as well as chapter 14 of this Volume: 'Statutory tests', Article 10 of the ECHR sets out the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 11 of the ECHR upholds the right to associate with others. We must exercise our duties under the Act in light of users' and services' Article 10 and 11 rights and not interfere with these rights unless we are satisfied that to do so is prescribed by law, pursues a legitimate aim, is proportionate to the legitimate aim and corresponds to a pressing social need.

8.108 We acknowledge that the safety defaults described in paragraph 8.48 for child users could impact the rights of both children and adults to freedom of expression and freedom of association. In particular:

- The settings restricting network expansion prompts and connection lists may make it harder for child users to make connections and communicate with other users. These settings restrictions may also make it harder for other users (both child and adult) to encounter and explore content produced by child users. These impacts may be particularly acute for child users with a public profile, such as those who wish to build a platform to share ideas or monetise their content.
- The measure's direct messaging and location information default settings may restrict legitimate communication and engagement between child users and other users. This impact may be particularly acute for children using U2U services where direct messages between users who are not connected is integral to the operation of key features or functionalities of the service (such as during certain gameplay scenarios). Where direct messages between users who are not connected are integral, we have mitigated the impact by recommending that services without a user connection functionality should provide a means for the child user to actively confirm they want to receive a direct message before receiving it.

8.109 However, overall, we believe the impact of our measure to be proportionate to the legitimate aim of reducing the risk of harm to children posed by perpetrators of grooming tactics, who may utilise the features restricted by our measure's default settings to make contract with potential grooming victims and to engage in behaviour that puts children at risk of CSEA offences. We also note that the impact would be mitigated by our recommendation that the measure's settings be implemented as safety defaults, so that child users may change them if they wish. Overall, we conclude that the measures contribute to the legitimate aim of preventing serious crime and protecting the health or morals of children. The measures are, therefore, a proportionate interference with the rights to the freedom of expression and association.

8.110 We did not receive any feedback from respondents to our November 2023 Consultation which disagreed with our assessment on the impact on human rights. However, we did receive concerns from a respondent about the potential negative impact of the safety defaults measure on adults engaging with child users for legitimate reasons.[1308] While we acknowledge that the measure could have some impact on users legitimately engaging with child users (as noted in paragraph 8.109), we consider this impact to be proportionate given the measure's legitimate aim of reducing grooming risks and child sexual exploitation and abuse. As discussed in 'Effectiveness', child users will also be able to manage their online experiences as the measure's settings are set to default. In other words, the settings can be 'turned off' by users if they wish, with children provided with a supportive message informing them of the risks of doing so.

8.111 Taking these points into account, we consider that the impact of the safety defaults measure on the right to freedom of expression and association to be limited and proportionate.

### Privacy and data protection

8.112 As explained in 'Introduction, our duties, and navigating the Statement', as well as chapter 14 of this Volume: 'Statutory tests', Article 8 of the ECHR sets out the right to respect for individuals' private and family life. An 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.

8.113 The UK General Data Protection Regulation and the Data Protection Act 2018 contain provisions intended to enhance the protection of children's personal data and the Children's Code provides standards that online services must follow when using children's data.[1309] We have considered these provisions and standards to ensure that children's data protection rights are not breached or disproportionately impacted by our measures.

8.114 We consider that there is no impact on the right to privacy associated with this measure. Some stakeholders raised concerns about the implications of the measure on both adults' and children's privacy. However, these concerns were not specific to a certain aspect of the measure. They were focused on the use of age assurance technologies affecting users' privacy and on potential conflicts with service providers' operability and ethos.[1310]

8.115 We note that service providers that are recommended to apply this measure will already have means to determine a user's age or age range and will likely already be collecting personal data for the purpose of these 'existing means'. Therefore, we expect that implementing our safety defaults measure, will not require providers to process more personal data than they currently do.

8.116 We acknowledge that some service providers may choose to adopt certain age assurance technologies or HEAA to target child users and implement this measure. We note that such technologies may require the increased processing of personal data and the increased collection of user data, which may impact the privacy of users. However, we note that our safety defaults measure does not require the use of age assurance for its implementation.

[1308] Name Withheld 3 response to November 2023 Consultation, p.17.

[1309] Set of standards for providers of online services using children's data implemented by the Information Commissioner's Office (ICO). ICO, 2022. Age Appropriate Design Code: a code of practice for online services [accessed 31 October 2024]. We refer to this as the 'Children's Code'.

[1310] Global Partners Digital response to November 2023 Illegal Harms Consultation p.20-21; Integrity Institute response to November 2023 Illegal Harms Consultation, p.16; Wikimedia Foundation response to November 2023 Illegal Harms Consultation, p.31-32.

Therefore, the provider's decision to introduce this measure to combat grooming and the decision to use age assurance or HEAA for their services are separate decisions. Where service providers do employ such technologies, we expect them to comply with relevant data protection legislation and to consider relevant guidance from the ICO. We have considered the case for recommending the use of HEAA, including the privacy implications thereof, separately in relation to measures for the proposed Children's Online Safety Code.[1311]

8.117 Furthermore, we expect that the measure will, in a number of respects, benefit child users in terms of privacy. It will add protection by restricting the visibility of child users' details on connection lists and network expansion prompts. On some services, it will also give child users more control of who they engage with by giving them the option to actively confirm acceptance before they interact with a user who they are not connected with, on a service without user connection functionalities.

8.118 We consider that the measure will improve children's control over their personal data (such as their name, photographs, and location), because this data will no longer be passively shared. Such data will be shared only with users with whom a child is already connected or with users who specifically search for them on a service.

## Who this measure applies to

8.119 In our November 2023 Consultation, we proposed that this measure would apply to:

- providers of U2U services that identify a high risk of grooming in their latest illegal content risk assessment; and
- providers of large U2U services that identify a medium risk of grooming in their latest illegal content risk assessment.[1312]

8.120 In addition, service providers should apply this measure based on two criteria.

- **The extent to which they possess the relevant functionalities**. Recommendations relating to setting defaults for network expansion prompts and connection list functionalities need only be applied by providers that have this functionality on their services.
- **The extent to which they can determine a user's age or age range**. Service providers should apply the measure to services with the existing means to determine a user's age or age range. This may be a form of age assurance or another method. While we are aware that some forms of determining a user's age or age range (such as self-declaration) are not as effective, the measure will still provide a significant degree of protection and potential benefits (see paragraphs 8.71 to 8.76).

### Assessment of who this measure should apply to

8.121 After considering stakeholder responses, our assessment of who this measure applies to remains unchanged. We conclude that this is proportionate considering the scale and

[1311] For more details, see our May 2024 Consultation. Ofcom, 2024. Protecting Children from Harms Online.

[1312] In the November 2023 Consultation, we also considered two other options regarding the types of providers who should apply this measure. Option 1: all large services identifying a high or medium risk of grooming; Option 2: (i) all services that have at least 25,000 child users AND identify a high risk of grooming and (ii) all large services identifying a medium risk of grooming. We detail our assessment of the two options in Annex 5.

severity of grooming, our assessment of the effectiveness of the measure, the costs to service providers of implementing it, and its impact on user rights.

8.122 Our evidence suggests that online grooming is a widespread and growing harm which can have a devastating impact on the lives of children.[1313] As explained in the 'Benefits' section (paragraphs 8.65 to 8.76), this measure will be effective in helping disrupt the grooming process and could bring about an important reduction in CSEA as a result. While it is not possible to precisely quantify the benefits that would flow from this measure, our evidence suggests that they could be very significant, given the harm these offences cause.[1314] The measure will have a number of direct and indirect costs. However, we consider these to be proportionate given the severity of the harm the measure is tackling and the scale and importance of the benefits of reducing this harm.

8.123 In the following sub-sections, we provide further information on our rationale for applying these measures to the in-scope services.

### Applying this measure to large services

8.124 For large service providers identifying a high risk of grooming, we consider the benefits from applying this measure in terms of reducing harms from online grooming are likely to be much greater than the costs. This is both because of the number of children that will likely benefit from the protection and because larger services will generally be well placed to absorb costs of the scale we have identified (even when we assume costs are at the top of the estimated range presented in paragraphs 8.89 to 8.95). We conclude that this is also likely to be the case for large service providers identifying a medium risk of grooming, given the number of children that use these services.

### Applying this measure to small services

8.125 The question of whether to extend the scope of the measure to incorporate smaller high-risk services is more finely balanced, given that fewer children will tend to use these services (relative to large services) and that small service providers may be less able to absorb the costs than large service providers.

### Applying this measure to small services assessed as high risk

8.126 As highlighted in the 'Benefits' section (paragraphs 8.65 to 8.70), our evidence shows that the functionalities targeted by this measure significantly increase the risk of grooming. Given the prevalence of grooming, services that offer these functionalities can pose a high risk of grooming even where they have relatively few child users. Restricting these functionalities through safety defaults could materially reduce this risk.

8.127 In the 'Benefits' section, we describe the severe and lifetime impacts of grooming on victims, which often extends to other children, communities, wider society, and public services. Given the severity of the harm, only a very small number of online grooming cases need to be averted for the measure to have a material benefit. We therefore consider that including small, high risk services in scope of the measure will confer significant benefits. The costs of the measure are likely to be towards the lower end of the estimated range for smaller service providers. This is because they will not have such high overhead and

[1313] See the Register of Risks chapter titled 'CSEA', particularly the section titled 'How CSEA offences manifest online' for a discussion of the scale and growth of these offences.

[1314] See the Register of Risks chapter titled 'CSEA', particularly the section titled 'Risk of harm to individuals presented by online CSEA offences'.

coordination costs (see paragraph 8.94). We therefore continue to consider it proportionate to apply the measure to small services whose risk assessment shows they pose a high risk of grooming.

8.128 In Annex 5: ‘Assumptions on costs and further analysis on costs and benefits’, we have undertaken a quantitative assessment of the direct costs of the measure compared with the benefits. On its face, the modelling we have done in this Annex would suggest that the benefits of the measure exceed the costs when it is applied to services with 25,000 or more child users, but that the benefits may not exceed the costs for services with significantly fewer child users. Notwithstanding the modelling in Annex 5, there are several reasons why we consider it appropriate and proportionate to apply the measure to all high-risk services regardless of size.

8.129 Firstly, as we explain in Annex 5, our model does not capture all of the benefits associated of the measure. We have only been able to quantify benefits of reducing contact CSEA. We have not been able to quantify the very significant benefits of preventing grooming which does not culminate in contact CSEA. Moreover, our estimate of the benefits of reducing contact CSEA is likely to be significantly understated as we have not been fully able to account for the long-term mental health impacts on victims and survivors.[1315] Nor has our model been able to quantify the impact of CSEA that results in death.

- We are also aware of the risk of displacement effects if we do not apply the measure to smaller services at high risk of grooming. If only the largest services implement this measure, this may result in perpetrators shifting to focus on targeting children on smaller services. As we show in the Register, we have observed displacement effects of this nature occur when large services have moved to improve protections against other harms.[1316] In light of the risks of displacement, we do not consider that we would be able effectively to achieve our policy objective of combatting online grooming if we excluded small but high risk services from the scope of the measure.
- The modelling does not factor in the role the measure may play in combatting other harms and the benefits that would flow from this.
- Stakeholder responses did not provide any compelling evidence to suggest that it would be disproportionate to apply the measure to very small high-risk services.

8.130 We therefore remain of the view that it is likely to be proportionate to apply this measure to services with small numbers of child users.

8.131 Respondents also raised similar concerns about the measure potentially displacing perpetrators onto smaller, less ‘risky’ services, or it causing younger users, due to social pressure, to be displaced onto services outside of scope of the measure.[1317]

[1315] Example of such deaths have been reported in the press. For example: Carrell, S. 2013. Scotland police investigate 'online blackmail' death of Fife teenager, *The Guardian*, 16 August; Dearden, L, 2018. Five British men have killed themselves after falling victim to online 'sextortion', police reveal. *The Independent*, 14 May; Campbell, J. and Kravarik, J., 2022. A 17-year-old boy died by suicide hours after being scammed. The FBI says it’s part of a troubling increase in ‘sextortion’ cases. *CNN*, 23 May; Yousif, N., 2022. Amanda Todd: Dutchman sentenced for fatal cyber-stalking. *BBC News*, 15 October. [All accessed 5 November 2024].

[1316] Register of Risks chapter titled ‘CSEA’.

[1317] Barnardo’s response to November 2023 Consultation, p. 20; Protection Group International response to November 2023, p. 10.

8.132 We expect that we have somewhat mitigated this risk due to who this measure applies to. In line with our risk assessment guidance, providers of smaller services which have risky functionalities most associated with grooming should assess as being at high risk for grooming, and hence the measure would apply to them.[1318] Moreover, our Risk Assessment Guidance ensures that services will be assessed as high risk for grooming, even where the targeted functionalities are not present, as long as there is evidence of grooming occurring on the service to a significant extent. Therefore, as there is a chance that perpetrators may be displaced to smaller services, we expect providers of such services to be alert to the possibility of such changes and update their risk assessment in line with our risk assessment guidance.

### Applying this measure to small services assessed as medium risk

8.133 In response to our November 2023 Consultation, we received suggestion from a stakeholder for broadening the scope of the measure to smaller services identifying a medium risk of grooming.[1319]

8.134 Our risk assessment guidance makes clear that providers of services that have the functionalities that put children most at risk of grooming (such as network expansion prompts and some connection lists functionalities) should assess as high risk of grooming. Our measure is recommended for all such services, regardless of size. The benefits of applying the measure to smaller services that are medium risk will be substantially lower than if they were high risk, because such services will not have some of the functionalities that put children most at risk. As opposed to large services with a medium risk of grooming, these smaller services will also tend to have fewer children accessing them. Therefore, any benefit from the measure is likely to be lower when applied to smaller medium risk services.

8.135 Considering our risk assessment guidance, and the lower impact that the measure would have on reducing grooming and the costs to providers of smaller services, we do not consider it proportionate to apply the measure to smaller services that assess as medium risk.

### Applying the measure to other kinds of illegal harm

8.136 In response to our November 2023 Consultation, some stakeholders suggested we apply the measure to all services that identify a medium to high risk of other kinds of illegal harms and not just those at risk of grooming.[1320] We have not done so as the measure is primarily designed to combat grooming. We note that it may have the ability to reduce other risks of illegal harms and have included the measure in the other duties Code alongside the CSEA Code.[1321]

## Conclusion

8.137 Our conclusion is that this measure will bring significant benefits to children's online experience by mitigating the risks of grooming for the purposes of CSEA on U2U services. This measure will impose some costs on both service providers and users. This includes

[1318] See 'Risk Assessment Guidance and Risk Profiles' for more details.
[1319] C3P response to November 2023 Consultation, p.23.
[1320] 5Rights Foundation response to November 2023 Consultation, p. 26; Molly Rose Foundation response to November 2023 Consultation, p. 34; OneID Ltd response to November 2023 Consultation, p.2.
[1321] See 'Benefits' section, paragraphs 8.74 to 8.76, for more details.

direct costs to service providers, alongside indirect costs to child users' experiences, such as making it more difficult to make friends online. Nevertheless, we consider the measure is proportionate, given the role of the functionalities outlined in facilitating grooming, and the severe impact that grooming has on children. We have therefore included within our Codes a recommendation that this measure should apply to:

- all providers of U2U services that identify a high risk of grooming in their latest illegal content risk assessment; and
- providers of large U2U services that identify a medium risk of grooming in their latest illegal content risk assessment.

8.138 In light of feedback surrounding, "an existing means of identifying child users", the final Codes wording specifies that service providers within the scope of this measure should have "an existing means of determining the age or age range of a particular user of the service concerned". This phrasing should clarify confusion surrounding this aspect of the measure.

8.139 We include this measure as part of the CSEA Code and Other Duties Code. We refer to it within these Codes as ICU F1.

## Measure on support for child users

8.140 In our November 2023 Consultation, we proposed that services that have existing means to identify child users and have particular functionalities provide child users[1322] with information at critical points in their online user journey to reduce the risks of grooming by enabling them to make informed choices. These are:

- When a child user seeks to change one or all of the recommended safety defaults (as outlined in our measure on safety defaults).
- At the point where a child user chooses to accept or deny a request from another user to establish a connection.
- At the point where a child user exchanges a direct message (either sent or received) with another user for the first time. Where direct messaging is a time critical element of a service functionality, information should be provided before that functionality begins.
- At the point where a child user takes action against another user's account (such as blocking, muting or reporting conduct).

8.141 We specified that this measure applies to all users aged under 18 and that the supportive information provided under this measure is prominently displayed and comprehensible for a child user.

8.142 We proposed that this measure should apply to providers of:

- U2U services at high risk of grooming; and
- large U2U services at medium risk of grooming.

[1322] The intention of this measure is for supportive information to be provided to users operating child user accounts, when the relevant user operating the child user account takes a particular action. However, for simplicity in our explanation of measures, we have referred to 'child users' when describing the effect and implementation of the measure in this chapter.

## Summary of stakeholder feedback[1323]

8.143 In response to the November 2023 Consultation, many respondents including civil society organisations, an online payments provider, and individuals, agreed with the proposed measure relating to support for child users.[1324] However, similar to the safety defaults measure, some of this support was caveated (as we explain in 8.147 and subsequent paragraphs).

8.144 Several civil society respondents noted their specific support for this measure alongside our safety defaults measure.[1325] 5Rights Foundation, Barnardo's and the Children's Commissioner for England, welcomed both measures.[1326] The NSPCC also commented that the measure is likely to "improve the efficacy of [the] new [safety default] settings and help children to make more informed decisions".[1327]

8.145 The UKSIC noted that both measures will not only "support the provision of supportive information in a timely and accessible manner to help users make informed choices when they seek to change their settings" but also support children's "wider digital literacy as well as provide potential protections against risks such as grooming and financial and online sextortion".[1328]

8.146 Our engagement with children also highlighted their support for the measure. They felt the measure was important to reduce the risk of children making uninformed choices when altering account settings.[1329]

8.147 Respondents also raised two main themes related to the measure. These were:

- effectiveness of the supportive information; and
- costs of the measure on child users' online experiences.

8.148 We examine these themes in the following paragraphs.

[1323] Note this list is not exhaustive, and further responses can be found in Annex 1: 'Further stakeholder responses'.

[1324] 5Rights Foundation response to November 2023 Consultation, pp.27; Barnardo's response to November 2023 Consultation, pp.19-21; Betting and Gaming Council response to November 2023 Consultation, pp.10; Children's Commissioner for England response to November 2023 Consultation, pp.22-23; Duran Dwyer's response to November 2023 Consultation, p.8; ICO response to November 2023 Consultation, pp.20; Nexus response to November 2023 Consultation, pp.14-15; NSPCC response to November 2023 Consultation, pp.36-37; One ID Ltd response to November 2023 Consultation, pp.2-3; Refuge response to November 2023 Illegal Harms Consultation, p.19; Segregated Payments Ltd response to November 2023 Consultation, pp.11; UKSIC response to November 2023 Consultation, pp.11; WeProtect Global Alliance response to November 2023 Consultation, pp.19-20.

[1325] 5Rights Foundation response to November 2023 Consultation, pp.27; Barnardo's response to November 2023 Consultation, pp.19-21; Children's Commissioner for England response to November 2023 Consultation, pp.22-23; NSPCC response to November 2023 Consultation, pp.36-37; UKSIC response to November 2023 Consultation, p.12.

[1326] 5Rights Foundation response to November 2023 Consultation, p.27; Barnardo's response to November 2023 Consultation, pp.19-21; Children's Commissioner for England response to November 2023 Consultation, pp.22-23.

[1327] NSPCC response to November 2023 Consultation, p.36.

[1328] UKSIC response to November 2023 Consultation, p.18.

[1329] Praesidio Safeguarding, 2024. Consulting children on proposed safety measures against online grooming. [accessed 16 December 2024].

## Feedback on the effectiveness of the measure

### “Age-appropriate” supportive information

8.149 Several respondents suggested that the format of the information should be “age-appropriate” so that messaging is understandable and effective for children in range of age groups.[1330] This theme included several points:[1331]

- 5Rights Foundation’s suggested the messaging should be “age-appropriate” and designed so that the information is “comprehensible” and “clearly presented”.[1332]
- BILETA was concerned that this measure (and the safety defaults measure) treated children as a homogenous group, and suggested separate messaging tailored for younger children (aged under 16) and older children (aged 16-18).[1333]
- Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) suggested that the messaging could be tailored to meet the needs of various age groups.[1334]
- [✂].[1335]
- Nexus commented on the need for the messaging to be “age-appropriate”, “factual” and “non-victim blaming”.[1336]
- [✂].[1337]
- WeProtect suggested the messaging should be child-friendly, accessible, and easy for all users to understand.[1338]
- INVIVIA suggested that this measure and measure ICU F1 provided “age-appropriate experiences”, noting that children of different ages have varying levels of maturity and risk tolerance.[1339]

8.150 Similar feedback was identified in both the November 2023 Consultation and our May 2024 Consultation. The NSPCC suggested that the provision of age-appropriate user support materials could be improved in the Illegal Harms and Protection of Children recommendations. It also called for Ofcom to produce guidance to help service providers

[1330] 5Rights Foundation response to November 2023 Consultation, p.27; BILETA response to November 2023 Consultation, p.14; CELE response to November 2023 Consultation, p.11; [✂]; INVIVIA response to November 2023 Consultation, p.19; [✂]; Nexus response to November 2023 Consultation, p.16; WeProtect response to November 2023 Consultation, p.20; Praesidio Safeguarding, 2024. [Consulting children on proposed safety measures against online grooming](). [accessed 16 December 2024].

[1331] We note that Yoti made a similar point in the May 2024 Consultation on Protecting Children from Harms Online, p.38, regarding service providers’ consideration of the provision of different versions of user support materials for different age groups of children. This feedback, while addressing a similar issue, has a different context as it is in relation to the measure we refer to as US6 (Provision of age-appropriate user support materials for children), which is not a directly comparable measure to measure ICU F2. For more details, see our May 2024 Consultation: Ofcom, May 2024. [Protecting Children from Harms Online]().

[1332] 5Rights Foundation response to November 2023 Consultation, p.27.

[1333] BILETA response to November 2023 Consultation, p.14.

[1334] CELE response to November 2023 Consultation, p.11.

[1335] [✂].

[1336] Nexus response to November 2023 Consultation, p.16.

[1337] [✂].

[1338]We Protect response to November 2023 Consultation, p.20.

[1339] INVIVIA response to November 2023 Consultation, p.19.

develop effective age-appropriate information.[1340] We respond to these themes in the ‘Effectiveness’ section (paragraphs 8.178 to 8.181).

### Consideration of users with disabilities or special educational needs (SEN)

8.151 Three respondents ([✂], Nexus and WeProtect) highlighted the need for the supportive information to be “disability friendly” and recognise users “with learning disabilities/communication difficulties”.[1341] We address these suggestions in the ‘Effectiveness’ section (paragraphs 8.182 to 8.185).

### Safety and privacy notices in supportive information

8.152 The NSPCC suggested that the benefits of safety and privacy should be included in the supportive information (particularly when children seek to change the safety defaults).[1342] It noted that safety and privacy features were not properly communicated to children in the measure (particularly when children seek to change or deactivate the default settings).[1343] We respond to this in the ‘Effectiveness’ section (paragraph 8.183).

### Persuasive design methods and broader presentation of the information

8.153 5Rights Foundation expressed concern about the use of supportive information in persuasive design strategies. It suggested the wording and presentation of the messaging could persuade child users into accepting lower standards of protection.[1344] It argued that we could provide greater specificity to service providers on the information contained within the messaging, which should include explicit warnings to children not to lower the protections offered by the safety defaults.[1345] Conversely, Snap argued for greater flexibility and discretion in how such information is presented to child users.[1346] We respond to this feedback in the ‘Effectiveness’ section (paragraphs 8.175 and 8.176).

[1340] NSPCC response to November 2023 Consultation p.37; NSPCC response to the May 2024 Consultation, p.69. We note that the NSPCC suggested in its May 2024 Consultation response, that the provision of supportive information under measure US6 (‘Provide age-appropriate user support materials for children’), could be improved in several areas, including by providing ‘age-appropriate’ supportive information, alongside developing messaging, which is “child-friendly”, “engaging” and is “non-victim” blaming. This feedback, while linked to the issue we discuss around safety and privacy notices, has a different context, as US6 is not a directly comparable measure to measure ICU F2 (support for child users). For more details, see our May 2024 Consultation: Ofcom, May 2024. Protecting Children from Harms Online.

[1341] [✂]; Nexus response to November 2023 Consultation, p.15; We Protect response to November 2023 Consultation, p.20.

[1342] NSPCC response to November 2023 Consultation, p.37.

[1343] NSPCC response to November 2023 Consultation, p.37.

[1344] 5Rights Foundation response to November 2023 Consultation, p.27.

[1345] 5Rights Foundation response to November 2023 Consultation, pp.27-28. We note a similar theme was raised by CP3 in its response to the May 2024 Consultation, p.31. It suggested that Ofcom should provide specificity around the nature and content related to US4 (Provision of information to child users when they restrict interactions with other accounts or content) and US5 (Signposting child users to support), including that the supportive information contained within the measures should include a defined minimum font size. However, we note US5 is not a directly comparable measure to measure ICU F2. For more details, see our consultation: Ofcom, May 2024. Protecting Children from Harms Online.

[1346] Snap response to November 2023 Consultation, p.21.

### Costs of the measure on child users' online experience

8.154 Snap raised concern that the measure could create an overly burdensome user experience, thereby creating an indirect cost to users.[1347] It suggested that the measure should contain fewer prompts and that the balance between information provided when a user makes a choice and information provided as part of the user onboarding experience should be taken into consideration. It also suggested that users should be given the option to dismiss future supportive information after displaying the information at the first instance of the user taking a triggering action.[1348] We provide our response to this feedback in the 'Costs and risks' section (paragraphs 8.194 to 8.196).

## Our decision

8.155 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation. We have made a slight amendment based on the feedback received, similar to our measure on safety defaults:

- In our final measure, we have adjusted the wording on which we consulted, replacing the phrase "an existing means of identifying child users" with "an existing means of determining the age or age range of a particular user of the service concerned" and have added a definition within the codes to explain our expectations of what such means may include. Given this is the same change as made in the safety defaults measure, we discuss the rationale behind it in the 'How this measure works' section under our safety defaults measure (ICU F1).
- We have made minor changes to the wording and definitions used in the measure. The changes clarify our expectations about how the measure should be implemented. Some changes include defining 'reporting conduct' (an action that can be taken against a user account), so providers are clear on what supportive information should be supplied to child user accounts. We discuss this particular clarification in the 'How this measure works' section below.
- We have also clarified that we consider this measure has the ability to mitigate other risks of illegal harms. We provide further details in 'The measure's ability to reduce other risks of harm' section (paragraphs 8.164 to 8.165).

8.156 The full wording of the measure can be found in our Illegal Content Codes of Practice for U2U Services, in which we refer to this measure as ICU F2.

## Our reasoning

### How this measure works

8.157 We recommend that all providers of U2U services with a high risk of grooming and all providers of large U2U services which identify a medium risk of grooming should offer

[1347] Snap response to November 2023 Consultation, p.21. We note that Snap raised a similar theme in response to the May 2024 Consultation, p.25 & p.27. This feedback however was in relation to measures we refer to as US1 (Notification for child users to actively confirm that they wish to be part of a group chat) and US5 (Signposting child users to support), which are not directly comparable measures to measure ICU F2 (support for child users). For more details, see our consultation: Ofcom, May 2024. Protecting Children from Harms Online.

[1348] Snap response to November 2023 Consultation, p.21.

supportive information to child users at four critical points (outlined in paragraph 8.159) in the user journey (where they have an existing means of determining who a child user is and where relevant functionalities exist).

8.158 This information should be prominently displayed and be clear and easy for a child user to understand. Services should consider their user base and ensure that the information is suitable for child users on their service.

8.159 The supportive information should be displayed at the point where:

- **A child seeks to change one of the recommended default settings as outlined in our safety defaults measure:** The information provided should assist children in understanding the implications of making this change, including the protections afforded by the default settings they are disabling. The information should provide beneficial friction (in terms of an opportunity to reflect on the impact of changing the default setting) for child users who may be disabling the settings because of pressure or blackmail.
- **A child makes a choice to accept or deny a request from another user to establish a connection:** The information provided should explain the types of interaction that would be enabled through establishing a connection, and how to take action against a user (such as blocking, muting, reporting conduct or equivalent actions). The provision of information, such as through blocking conduct, should help equip children with knowledge of how to protect themselves, in the even that any future engagement with that user cause them to feel uncomfortable or unsafe.
- **A child user exchanges a direct message (either sent or received) with another user for the first time:** The information provided should remind the child that this is the first direct communication with that user and explain how to take action against that user. These messages should explain the relevant risks associated with communicating through a direct message functionality with users they do not know. The information should support the child user to pause before engaging with a new user and support the child in stopping or disrupting engagement with a user they communicate with through direct messaging.
- **A child user takes action against another user's account (such as blocking, muting, or reporting conduct):** The information provided should support the child user to understand the effect of the action (including the types of interactions it would restrict and whether the user would be notified) and indicate the further options available to limit interaction. The information should also point to actions that child users can take to increase their safety, which can include pointing to safety and privacy settings that are available to the child user on the service. In relation to reporting conduct, the action relates to the safety duty to operate a complaints procedure and ensure it is easy for children to access and use. In particular, providers should support children in making complaints about behaviours or actions taken by the user associated with the relevant user account that may lead to a potential breach of terms and conditions of the service, or that may affect the ability of the provider to comply with their safety duties under section 10 of the Act.

8.160 We have identified these stages as being key points at which supportive information would be particularly effective at mitigating the risk of harm to child users.[1349] We consider this will provide children with timely information so that they can make informed choices about their safety in their online experiences.

8.161 In the 'Benefits' and 'Effectiveness' sections, we further discuss why we consider that the provision of information at these crucial points would be particularly effective at mitigating and managing the risk of harm to child users online.

## Benefits

8.162 We consider there are significant benefits that can arise from providing information to child users at critical points during their use of a service. Broadly speaking, there are three aspects to the benefits that arise from the provision of this information.

a) **Information leading to reassessment of a user choice:** for example, a child may decide not to turn off a default setting after being provided with information that informs them of the potential risks involved, and so will be safer.
b) **Information leading to increased awareness of the potential risks when interacting with other users online:** for example, a child may accept a message from an unknown user but do so with greater understanding of how to take action against that user if they feel uncomfortable, and so will be safer during such interactions.
c) **Information that leads to a child's increased knowledge in online safety:** for example, individual child users are likely to feel more informed on risks to their online safety with the information provided. They are also likely to be more aware of the tools to help them mitigate these risks.

8.163 In sum, we therefore consider that the measure will make children less susceptible to grooming, thus delivering significant benefits. We expect the measure will reduce the grooming risks and bring about an important reduction in the sexual abuse of children. We note that several stakeholders also supported the measure, albeit some with caveats.[1350] [1351]

### The measure's ability to reduce other risks of harm

8.164 We consider that this measure will help child users make more informed choices on the use of the functionalities, which have been identified in the Register to be of particular risk to children. We consider that this measure will therefore assist providers in mitigating risks of harm related to other offences, because of the support it provides to a child user at the point of changing the safety settings and the support it provides child users around the use of risky functionalities.

[1349] See our November 2023 Illegal Harms Consultation for further details: Ofcom, 2023. Protecting people from Illegal Harms Online, pp.255-260.

[1350] We note stakeholders' feedback in relation to this measure, in 'Summary of stakeholder feedback' section, and address this throughout the chapter.

[1351] 5Rights Foundation response to November 2023 Consultation, p.27; Barnardo's response to November 2023 Consultation, p.19-21; Betting and Gaming Council response to November 2023 Consultation, p.10; Children's Commissioner for England response to November 2023 Consultation, p 22-23; Duran Dwyer's response to November 2023 Consultation, p.8; ICO response to November 2023 Consultation, p.20; Nexus response to November 2023 Consultation, p.15; NSPCC response to November 2023 Consultation, p.36; One ID Ltd response to November 2023 Consultation, pp.2-3; Refuge response to November 2023 Consultation, p.19; Segregated Payments Ltd response to November 2023 Consultation, p.11; UKSIC response to November 2023 Consultation, p.17; WeProtect Global Alliance response to November 2023 Consultation, pp.19-20.

8.165 In the 'Benefits' section (paragraphs 8.72 to 8.73) for our safety defaults measure, we explained the impact of perpetrators using functionalities to target children. As the supportive information provided through this measure can directly influence the use of the safety settings, we consider this measure will mitigate the same risks of harm related to other offences that the safety defaults measure mitigates.

## Effectiveness

8.166 In this section, we provide more details on the effectiveness of the measure, including why we think providing supportive information at key points in the user journey would be effective.

8.167 As we discussed in our November 2023 Consultation, prompts can be used to influence user behaviour. In this case, they can be used to improve a user's safety while at the same time preserving the option for the user to choose another course of action. We also consider that the timing of the supportive information will help children make more informed choices regarding their safety, by giving them the information that they need at the right time.

8.168 The need to factor in considerations of timing and relevance is also consistent with what has been advocated by others.

- In their 'Safety by Design' principles, the Australian eSafety Commissioner recommend leveraging "the use of technical features to mitigate against risks and harms, which can be flagged to users at point of relevance, and which prompt and optimise safer interactions."[1352]

- In their responses to the 2022 Illegal Harms call for evidence, 5Rights Foundation described the desirability of "just-in time-warnings, informing users of potential risks associated with content they are about to interact with," which was echoed as an effective strategy to mitigate risk of illegal harm in examples of current practice cited by both the Alan Turing Institute and Glitch.[1353]

- In an article for Trust, Transparency & Control (TTC) Labs, Meta data strategist Dr Dan Hayden highlighted the importance of giving the user "the right information, at the right time" – in other words, when it becomes "relevant to the action the user wants to take".[1354]

8.169 Academic and regulatory work has suggested that prompts can influence user in general to make safer choices.[1355] There is also some research which is specific to children's online

[1352] Australian e-Safety Commissioner, 2019. Safety By Design Principles and Background. Principle 2.3 [accessed 1 November 2024].

[1353] 5Rights Foundation response to 2022 Ofcom Call for Evidence: Second phase of online safety regulation, p.27; Alan Turing Institute response to 2022 Ofcom Call for Evidence: First phase of online safety regulation, pp.8-9; Glitch response to 2022 Ofcom Call for Evidence: First phase of online safety regulation, p.9.

[1354] TTC Labs (Hayden, D.), 2021. Making Sense of Data Disclosures: Leveraging Context in Design. [accessed 1 November 2024].

[1355] For example: European Commission, 2019. Study on media literacy and online empowerment issues raised by algorithm-driven media services. [accessed 25 November 2024]; US Food and Drug Administration, 2019. Communicating Risks and Benefits: An Evidence-Based User's Guide. [accessed 1 November 2024]; Ioannou, A., Tussyadiah I., Miller G., Li S. and Weick M., 2021. Privacy nudges for disclosure of personal information: A systematic literature review and meta-analysis. *PLoS One*, 16 (8). [accessed 25 November 2024]; Acquisti et al., 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Computing Surveys,* 50 (3). [accessed 25 November 2024].

behaviour. Research on mechanisms for enhancing the privacy risk awareness of teenagers online indicates that the characteristics of children and teenagers put them at greater risk of harm (as they tend to be "trusting, naïve, curious, adventuresome, and eager for attention and affection"). However, this research also finds that using nudges in the form of prompts can influence teenagers' decision-making without restricting their freedom to decide: the characteristics of children may make them vulnerable to harm, but the prompts will make them more aware of their options without directly trying to restrict their natural behaviours. [1356] This evidence supports the argument for seeking to protect children through the use of prompts, or supportive information.

### Providing support at the four critical points in the user journey

8.170 The provision of information at critical points in the user journey will help children make more informed choices regarding their safety by giving them the information they need at the right time. In the following section, we outline the critical points at which we think supportive information will be particularly effective in mitigating and managing the risk of harm.

- **When a child seeks to turn off one of the recommended safety defaults:** We consider there to be residual risks associated with the safety defaults proposals in respect of network expansion prompts, direct messaging, and location functionalities. As outlined, our recommendations are choice-preserving. This means that child users can choose to disable the default settings or change the settings to less privacy and safety enhancing settings, which may reintroduce the risks that the safety defaults measure is designed to address. Child users may choose to disable a default for many positive reasons, including finding new connections or followers, increasing the reach of content they create, or wanting to have more features or functionalities available to them. Children may also wish to set their functionality settings to emulate adults. However, they may also experience pressure to turn off safety defaults from other service users, including perpetrators of grooming. We are concerned about instances of child users seeking to disable these default settings without fully understanding the relevant grooming risks and other kinds of illegal harms risks they could be introducing. The information provided should therefore assist children in understanding the implications of making this change, including reiteration of the protections afforded by the default setting they are disabling. Once informed of these risks, a child may change their mind about moving away from the default safety settings. The prompt together with the information it contains will also act as 'positive' friction. It interrupts the user journey and provides an opportunity to reflect on changing the default setting, particularly for child uses who may be disabling settings because of pressure or blackmail. In light of the evidence outlined in paragraphs 8.168 to 8.169 on the user safety benefits of prompts, we consider that providing support at this stage of the user journey would aid in reducing the risk of harm to child users.
- **At the point where a child is making a choice to accept or deny a request from another user to establish a connection:** We expect that children will continue to receive and accept friend requests, including from people they may not know. While research suggests that there are benefits to children connecting with other users online, this also

[1356] Alemany, J., del Val E., Alberola, J., and García-Fornes, A., 2019. Enhancing the privacy risk awareness of teenagers in online social networks through soft-paternalism mechanisms. *International Journal of Human Computer Studies*, 129. [accessed 1 November 2024].

poses a risk as it can also increase the likelihood of children accepting and connecting with potential perpetrators (as outlined in the 'CSEA' chapter of the Register, specifically the section 'Grooming'). The information provided should explain the types of interaction that would be enabled through establishing a connection, and should give details on how to take action against a user. We anticipate that these informational prompts may provide a pause before a formal connection is established with another user. Slowing down this decision-making could disrupt early engagement with perpetrators in the grooming journey where perpetrators seek to establish contact with child users.

- **At the point where a child exchanges a direct message (either sent or received) with another user for the first time:** We expect that children are likely to continue to receive direct messages from new connections that they have made online, which presents a risk in circumstances where that person is seeking to groom or otherwise sexually exploit the child (as outlined in the 'CSEA' chapter of the Register, specifically the section 'Grooming'). As explained in paragraph 8.73, direct messaging functionalities can be exploited by perpetrators as a means of initiating the grooming process through private communications. Although children may be aware of the risks of harms when interacting with people online, they are often unsure how to avoid them.[1357] The information provided should remind the child that this is the first direct communication with that user and explain how to take action against that user. This informational prompt should provide the child user with a pause before interacting with a user and cause the child to consider their engagement with a new user.[1358] We also consider that children knowing how to take action against users will empower them to protect themselves should their future engagement with a specific user become uncomfortable or unsafe.

- **At the point where a child user is taking action against another account, including blocking, muting or reporting conduct:** Research suggests that children find it difficult to end contact with perpetrators of grooming or other offences online – sometimes as a result of blackmailing or threats they are experiencing.[1359] This is despite children being more likely to use online reporting tools compared to turning to offline support systems, such as a caregiver or friend.[1360] Research also indicates that children are more likely to block users than report them as they may feel unclear about the process, which can discourage reporting.[1361] Furthermore, ending or restricting contact during online grooming is a distressing time for a child for a variety of reasons. Some children may worry about the repercussions if a perpetrator finds out they have taken action against

[1357] Macaulay J.R., Boulton, M., Betts, L., Boulton, L., Camerone, E., Down, J., Hughes, J., Kirkbride, C. and Kirkham, R., 2019. Subjective versus objective knowledge of online safety/dangers as predictors of children's perceived online safety and attitudes towards e-safety education in the United Kingdom. *Journal of Children and Media*, 14 (3). [accessed 15 October 2024].

[1358] We consider that it would be appropriate to allow an alternative approach to providing this information in circumstances where, on a particular service, receiving a direct message is a necessary and time-critical element of another service functionality that a child user is engaging with. In that case, the child user may be provided with this information before any interaction associated with the functionality begins.

[1359] Hanson, E., 2017. The Impact of Online Sexual Abuse on Children and Young People: Impact, Protection and Prevention, in Brown, J (ed.) *Online risk to children: Impact, protection and prevention*. Oxford: Wiley Blackwell/NSPCC, pp. 97-122. [accessed 15 October 2024].

[1360] Thorn, 2021. Responding to Online Threats: perspectives on Disclosing, Reporting, and Blocking. [accessed 15 October 2024].

[1361] ibid.

them.[1362] These threats and coercive tactics can have significant impact on the child's mental health, which can include self-blame, negative sense of self, depression, anxiety and suicidal ideation (which includes reports of victims and survivors taking their own lives following incidences of grooming and perpetrators threatening to share their sexual images).[1363] The information provided should help the child user understand the effect of the action (including the types of interactions it would restrict and whether the user would be notified) and indicate the further options available to limit interaction and increase their safety. For example, the information could encourage children to report users for actions or behaviours in breach of a service's terms and conditions or empower them to change either their safety or privacy settings on their account (thereby preventing perpetrators from sending messages from other accounts). We consider that the provision of this information will be particularly beneficial to child users who have had a negative experience and who require additional knowledge to make informed decisions to support their ongoing safety on a service. This can also have positive ramifications for child users feeling safer online after having taken action on the service. We also note that service providers have a duty to operate an easy-to-use and transparent complaints procedure for when a user (including a child user) feels that the provider is not complying with its illegal content safety duties. This includes complaints against users who may be increasing the risk of harm to individuals. Therefore, we consider that further information about the reporting process (either as an action taken or potential next step) will provide clarity for child users and may increase confidence in the reporting process.

8.171 In light of this analysis, we consider that providing relevant information at these four critical points would be particularly effective at mitigating and managing the risk of grooming harm to child users online. This reinforces our view that the measure under consideration would deliver significant benefits. As discussed above, we also consider that this measure supports the effectiveness of the safety defaults measure by increasing the likelihood of them not being switched off.

### Format of supportive information

8.172 We recognise that the format of the supportive information is likely to determine their effectiveness. However, as we noted in our November 2023 Consultation, the available evidence regarding how to present user support messages does not point to a single 'best practice' approach.

8.173 In the several studies that found similar mitigations to be effective, the authors pointed to the various factors that they considered to contribute to the effectiveness of using nudges, such as length, colour, and language. However, there was no consistent recommendation

[1362] Joleby, M., Lunde, C., Landstrom, S., and Jonsson, L. S. 2020. "All of me is completely different": Experiences and consequences among victims of technology-assisted child sexual abuse. *Frontiers in Psychology*, 11, 606218. [accessed 07 November 2024].

[1363] Example of such deaths have been reported in the press. For example: Carrell, S. 2013. Scotland police investigate 'online blackmail' death of Fife teenager, *The Guardian*, 16 August; Dearden, L, 2018. Five British men have killed themselves after falling victim to online 'sextortion', police reveal. *The Independent*, 14 May; Campbell, J. and Kravarik, J., 2022. A 17-year-old boy died by suicide hours after being scammed. The FBI says it's part of a troubling increase in 'sextortion' cases. *CNN*, 23 May; Yousif, N., 2022. Amanda Todd: Dutchman sentenced for fatal cyber-stalking. *BBC News*, 15 October. [All accessed 5 November 2024].

on how to apply these factors.[1364] In its Children's Code, the ICO says providers "should bear in mind children's needs and maturity will differ according to their age and development stage" and provides a guide for considering the interests, needs, and evolving capacity of children at different ages.[1365]

8.174 We noted some variation in the technical interfaces through which services communicate safety nudges to users. Some present this information as a pop-up, while others embed it within an interface. However, we concluded that we did not have sufficient evidence to understand any differences in the effectiveness of these approaches.

8.175 We further note 5Rights Foundation's concern about supportive prompts' use in persuasive design strategies, which they felt could lead to children accepting lower standards of protection.[1366] We were not presented with substantive evidence concerning the design strategies that service providers may use to present the prompts (including evidence that providers may attempt to persuade children into accepting lower standards of protection). We therefore do not consider it necessary to change our approach at this stage. If presented with such evidence in the future, we will consider it accordingly.

8.176 We consider that we have addressed Snap's concern regarding service providers' discretion in designing and presenting supportive information, as the measure does not prescribe a specific format for the wording or design of the support prompts.[1367]

8.177 Given the lack of a one size fits all approach, we suggested that services are best placed to design, test, and evaluate the format and delivery of the supportive information to optimise the benefits for the child users on their services. We did not make specific recommendations around how the supportive information should be presented and encouraged services to establish their own best practice on how to deliver information to child user accounts on their services.[1368] However, we did specify that service providers should provide children with clear, comprehensible and easy to understand information at critical points in their user journey (as discussed in 'Our reasoning' section). We have also set out our recommended approach on the nature of the information that should be provided and asked providers to consider their service's user base when designing both the information and its delivery. This will enable child users to make informed choices about risk in their online experiences.

### "Age-appropriate" supportive information

8.178 Several respondents suggested that the measure could be made more effective by considering the needs of different age groups. They suggested that the format of the supportive information should be "age-appropriate", so that messaging is understandable and effective for children in a range of age groups.[1369] The NSPCC also stated that it would

[1364] For example, Ioannou, A. et al., 2021, 1. This literature review called for further research "to elucidate the relative effectiveness of different intervention strategies and how nudges can confound one another".
[1365] ICO, 2022. Age Appropriate Design Code, p.31. [accessed 15 October 2024].
[1366] 5Rights Foundation response to November 2023 Consultation, p.27.
[1367] Snap response to November 2023 Consultation, p.21.
[1368] Ofcom's Behavioural Insights Hub has published its own research on the use of prompts to improve user engagement with safety measures. It would welcome the opportunity to discuss ways of designing, testing and evaluating the effectiveness of different forms of support messaging with service providers.
[1369] 5Rights Foundation response to November 2023 Consultation, p.27; BILETA response to November 2023 Consultation, p.14; CELE response to November 2023 Consultation, p.11; [✂]; INVIVIA response to November

be useful to providers to have specific guidance on the development of effective age-appropriate information for measures in the proposed Illegal Content and Children's Online Safety Codes and suggested such guidance could be supplied by Ofcom.[1370]

8.179 We recognise that providing age-appropriate supportive information for children supports their online safety. Informed children are better able to take appropriate action when something goes wrong online.[1371] This is particularly important if children are repeatedly exposed to harms online, and for children who might not have access to adults who can help them stay safe online.

8.180 We expect service providers to consider the needs of different age groups as part of the development of supportive information. We have stipulated that the information should be presented in a way that is clear, comprehensible, and easy for all children to understand. It should also be displayed prominently to children at relevant critical points in the user journey. The requirements of the measure therefore ensure that children of all ages can understand the information and benefit from the protections that they provide.

8.181 We do not consider it appropriate at this time to provide service providers with specific guidance on how to make the supportive information 'age appropriate' as there is no clear consensus on what good practice looks like and the range of services in scope of the measure means a 'one size fits all' approach is unlikely to yield good outcomes.

### Providing specificity for users with specific needs

8.182 We also acknowledge the suggestions from respondents to provide guidance or increase the level of detail we prescribe for the formatting of the information, or to provide specificity for users with specific needs (including users with disabilities or SEN).[1372]

8.183 Based on our assessment of the evidence and information we currently have, our position on the format of the supportive information remains unchanged following the November 2023 Consultation. Giving service providers flexibility, enables them to develop wording that they see fit for their users. This includes the ability to develop specific wording or the presentation of information for children with disabilities or SEN, or to consider additional safety and privacy notices for children.

8.184 We are not prescriptive about how the information should be formatted and provided to child user accounts. Therefore, at this stage, we do not think it appropriate to specify particular ways in which it should be accessible to disabled people. We would expect approaches to accessibility to vary from service to service, subject to their service's features and design, and on that basis recommend providers are best placed to decide how to ensure information is accessible to disabled people. However, providers should consider

2023 Consultation, p.19; [✂]; Nexus response to November 2023 Consultation, p.16; WeProtect response to November 2023 Consultation, p.20

[1370] NSPCC response to May 2024 Consultation, p.69.

[1371] Ofcom, 2022. Serious game pilot: Trialling a serious game as an approach to making children safer online. Subsequent references are to this research. Note: All participants (n = 629) were aged between 13 and 17.

[1372] [✂]; Nexus response to November 2023 Consultation, p.15; NSPCC response to November 2023 Consultation, p.37; WeProtect response to November 2023 Consultation, p.20.

their obligations under other relevant legislation (for example, the Equality Act 2010) and, where relevant, appropriate guidance.[1373]

8.185 In summary, we consider that providing supportive information at the critical points identified above, would be particularly effective at mitigating and managing the risk of harm to child users online. The measure will likely contribute in an important way to the reduction of grooming and child sexual exploitation and abuse online. Given the severity of the impact of grooming and sexual abuse and the prevalence of these harms, we consider that the benefits flowing from the measure would be significant.

## Costs and risks

8.186 In our November 2023 Consultation, we considered the direct costs to service providers in implementing this measure. We also considered the indirect costs to children and to service providers.

8.187 Our assessment of these costs, after considering stakeholder responses, remain broadly unchanged from the November 2023 Consultation.[1374]

### Direct costs

8.188 Direct costs for service providers are likely to be largely one-off costs. They would consist of developing information to present, and system changes to implement its appearance at the four critical points. There would also be some on-going costs to maintain the functionalities. Since we are not proposing to prescribe precisely how services should provide information (see 'Format of supportive information' section), we expect there to be a range of costs depending on how much development is put into crafting the way information is provided and the functions needed to provide said information.

8.189 We estimate the one-off upfront direct cost to be approximately £30,000 to £325,000, made up of six to 36 months' worth of staff resources.[1375]

8.190 The variation in costs is driven by differences in both the size of a service and of the provider's existing systems. Providers with larger and more sophisticated services may incur higher costs due to iterative testing and evaluation of supportive information formats and delivery methods. Material costs may also be incurred if a service provider does not already have a system in place to provide supportive information. For smaller services, we would expect the costs would tend to be towards the lower end of the range. This is because

[1373] See for example, 'WCAG 2 Overview', 2005. World Wide Web Consortium's (W3C) Web Content Accessibility Guidelines (WCAG), *W3C Web Accessibility Initiative (WAI)*, March 2024. [accessed 15 November 2024].

[1374] We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS. However, since our cost estimates are rounded, these changes do not always result in the rounded estimate changing. We received some general feedback on the cost assumptions (such as salary assumptions) that are fed into these costs. We consider that feedback and our updated wage assumptions in Annex 5.

[1375] We estimate the one-off upfront engineering cost associated with the implementation of all recommended supportive information under the measure to be approximately £30,000 to £115,000, made up of six to 12 months' worth of staff resources (split equally between software engineering staff and other professional occupation staffs such as project managers). We further estimate the one-off upfront overhead and coordination costs associated with any change to the frontend or backend systems of a service to be approximately £0 to approximately £210,000, made up of zero to 24 months' worth of professional occupation staff resources. All estimated are based on our standard assumptions for labour costs set out in Annex 5.

providers of smaller services will tend to have lower overhead and coordination costs in making changes (see paragraph 8.94).

8.191 In addition to the upfront costs, we assume the ongoing costs to review and monitor the measure are 25% of the upfront costs on an annual basis, ranging from £7,500 to £81,250.[1376]

### Indirect costs

8.192 We also consider indirect costs to service providers which may result from lower user engagement (and thus potentially lower advertising revenue) on their service. While this may result in lost revenue to providers, we consider this is likely to be proportionate given we think that the measure will materially reduce the risk of grooming online.

8.193 We consider indirect costs to child users on a service who may spend time and effort engaging with the information provided under this measure. We conclude these costs to be relatively small as the information points are not designed to be frequent or intrusive (see the 'Effectiveness' section). Also, providers do not have to require users to confirm that they have read or seen such information. This is likely to make the measure less intrusive and should reduce the disruption to the user's experience.

8.194 In response to our November 2023 Consultation, Snap disagreed with our assessment and expressed concern that this measure could create an overly burdensome user experience.[1377]

8.195 We acknowledge concerns regarding the burden that the measure may create for children and recognise that some research indicates prompts, while effective, can also be perceived as "annoying".[1378] Frequent exposure to pop-up alerts can also lead to users becoming de-sensitised to the warning contained in the alert.[1379] This means that the timing and relevance of such interventions becomes particularly important in ensuring that they achieve the desired effect.

8.196 Having considered available evidence (as discussed in the 'Benefits' and 'Effectiveness' sections), we conclude that supportive messages will materially improve children's safety online. As the user journey points at which supportive information is recommended under the measure are not expected to occur frequently and will happen only at certain critical intervals, we consider the burden on child users to be proportionate. We also consider that the indirect costs of the measure may be reduced by the fact that services have flexibility in how they design the supportive information. In other words, they will have scope to implement them in such a way which minimises unnecessary interference with child users' experiences.

[1376] Based on our standard assumptions for ongoing maintenance of software changes set out in Annex 5.

[1377] Snap response to November 2023 Consultation, p.21; Snap response to May 2024 Protecting Children from Harms Online Consultation, p. 25 & p.27.

[1378] Micallef, N., Just, M., Baillie, L., and Alharby, M., 2017. Stop annoying me!: an empirical investigation of the usability of app privacy notifications. *Association for Computing Machinery*. Proceedings of the 29th Australian Conference on Computer-Human Interaction. [accessed 15 October 2024].

[1379] Bravo-Lilo, C., Cranor, L., Komanduri, S., Schechter, S., and Sleeper, M. (2014). Harder to Ignore? Revisiting Pop-Up Fatigue and Approaches to Prevent It. *Symposium on Usable Privacy and Security (SOUPS)*, July 9-11, Memlo Park, CA.

## Rights impact

### Freedom of expression and freedom of association

8.197 We explained in paragraphs 8.107 to 8.111 the right to freedom of expression and the right to associate with others, as well as Ofcom's duties in respect of these rights. We have considered these rights in respect of this measure and consider that the measure may have a limited impact on the rights to freedom of expression and freedom of association.

8.198 We recognise that the measure introduces friction to child users' ability to turn off default settings and that such friction may delay or restrict child users' ability to connect and communicate with other users. Such a delay could influence their behaviour and may result in them being less likely to establish new connections or communicate with new users online. However, we expect any such delay to have a minor impact on child users' rights to freedom of expression and association overall as they will not be prevented from turning off the default settings nor will they be prevented from adding or communicating with new connections, if they wish to do so.

8.199 We also expect the measure to have a minor impact on adult users' rights to impart information and ideas to child users. This is because our measure does not restrict the adult user from sharing information and ideas intended for children altogether – and if a child wants to receive information and ideas targeted at them, they can turn 'off' the default settings.

8.200 We consider that the above impacts on child users' and other users' rights of expression and association do not amount to an undue interference. We consider that the abovementioned minor impacts on rights are proportionate to the measure's legitimate aim of reducing the risk of harm, which is achieved by increasing children's awareness of the risk associated with certain activities on a service.

### Privacy and data protection

8.201 We explained in paragraphs 8.112 to 8.114 the right to privacy and Ofcom's duty in respect of these rights. We also explained in paragraph 8.113 the relevant data protection legislation and guidance available that we have considered alongside the right to privacy in respect of this measure.

8.202 We consider that this measure will have no impact on children's rights to privacy and will have a minimal impact on children's data protection rights. While delivery of a supportive message may require extra processing of the child's personal data, it will not have a significant negative impact on the child's right to privacy. On the contrary, this measure to warns them of the consequences of their actions and will give child users agency over their own privacy. There is no expectation from this measure for service providers to extract or retain information relating to a child user's engagement with the supportive information provided or the actions taken in relation to such information (beyond that which would be extracted or retained during the normal running of the service or for the action of changing a setting).

## Who this measure applies to

8.203 We recommend this measure for all service providers that fall within the scope of the safety defaults measure. These include:

- all providers of U2U services that identify a high risk of grooming in their latest illegal content risk assessment; and

- providers of large U2U services that identify a medium risk of grooming in their latest illegal content risk assessment.

8.204 As with the safety defaults measure, this measure applies to providers that have an existing means to determine a user's age or age range and where relevant functionalities exist.

8.205 Our rationale for applying the measure to these services is the same as the explanation provided in the 'Who this measure applies to' section under the safety defaults measure.

## Other factors in the development of the measure

8.206 The following section provides an overview of other issues that have influenced our final supportive information measure (beyond the stakeholder feedback discussed in the preceding sections).

### Amending supportive information to include both accounts and content

8.207 In the November 2023 Consultation, we proposed that supportive information should be provided to children when they have taken action against another user (such as blocking, muting, or reporting conduct). We recommended that this information should include details on the effect of the action (such as the interaction that would be restricted), whether the user in question would be notified, and further options available to limit interaction or increase user safety.[1380]

8.208 However, a similar measure (PCU E2) proposed in the May 2024 Consultation recommends providing supportive information to children when they take restrictive action against another user account **or content**.[1381] This should include information about the effect of the action and other actions child users may take to protect themselves further. As with the supportive information measure in this chapter, this could include information on reporting, blocking or muting tools, among others.

8.209 In our May 2024 Consultation, we noted that we would consider whether to change the supportive information measure discussed in this chapter (measure ICU F2) to include information that should be provided when a child user takes action against content, in addition to against user accounts. We are continuing to actively consider this extension. If we do decide to change the measure, we will consult on proposed amendments.

# Conclusion

8.210 In light of our analysis, we consider that there are significant benefits to be gained from the provision of supportive information to child users at critical points during their user journey. By increasing the availability of this information to children, this measure will help protect them from the risk of grooming and will improve the effectiveness of our safety defaults measure. While this support to child users measure will impose some costs, we consider that these would be proportionate given the severity of the harm caused by grooming and the role the supportive information will play in combatting that harm.

[1380] Ofcom, November 2023. 'Protecting people from illegal harms online', p.230.

[1381] This proposed measure's scope differs from that of the measures in this chapter. Measure PCU E2 is recommended for large U2U services that are multi-risk for content harmful to children with supportive information when they take action against another user or kind of content (see: Ofcom, May 2024. 'Protecting children from harms online', *Chapter 14. Developing the Children's Safety Codes: Our framework*, Vol 5: What should services do to mitigate risks?, pp.386-387).

8.211 Our final measure will apply to:

- all providers of U2U services that identify a high risk of grooming in their latest illegal content risk assessment; and
- providers of large U2U services that identify a medium risk of grooming in their latest illegal content risk assessment.

8.212 As with the safety defaults measure, this support for child users measure applies where services have an existing means to determine the age or age range of a user and where relevant functionalities exist.

8.213 We will include this measure as part of the CSEA Code and other duties Illegal Content Codes (as set out in 'Illegal Content Codes of Practice for U2U services'), within which this measure is referred to as ICU F2.

# 9. Search settings, functionalities and user support

## What is this chapter about?

Search services can act as a gateway to illegal content that exists online. In particular, search functionalities and features that have been designed to optimise search results can inadvertently make it easier for users to encounter illegal content in those results.

There are steps that service providers can take to reduce these risks and make it less likely that users encounter illegal content through their service. In addition, providers can offer supportive information to users to allow them greater choice and control over their experiences on search services. This will help to reduce the likelihood that users seek out illegal content and mitigate the risk of harm they may face as a result.

This chapter set out and explains the rationale for a series of measures we are recommending providers of search services take to protect people from illegal content, and to which search services they should apply.

## What decisions have we made?

We are recommending the following measures:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| ICS F1 | Providers should offer users a means to **easily report predictive search suggestions** which they believe can **direct users towards priority illegal content**. If a **clear and material risk** is identified, the provider should take appropriate steps to ensure that the reported predictive search suggestion **is not recommended to any users.** | Providers of large general search services that use a predictive search functionality |
| ICS F2 | Providers should detect and **provide warnings and support resources** in response to **search requests** where the wording clearly indicates that the user may be seeking to encounter **child sexual abuse material** (CSAM). | Providers of large general search services |
| ICS F3 | Providers should provide **crisis prevention information** in response to **search requests** that contain **general queries regarding suicide** and queries seeking **specific, practical, or instructive information regarding suicide methods.** | Providers of large general search services |

## Why are we making these decisions?

Our first measure (ICS F1) will reduce barriers to reporting predictive search suggestions that can direct users to encounter priority illegal content. This will raise providers' awareness of problematic search suggestions and enable them to ensure that they are no longer recommended to users. In doing so, this measure will reduce the likelihood of users being prompted to run those searches, and encountering illegal content as a result.

Our second measure (ICS F2), on CSAM warning messages, is designed to deter potential perpetrators from accessing CSAM via the search results by providing them with resources that may help them refrain from committing CSEA offences. By reducing searches for CSAM, the measure may also reduce the harm inflicted on child victims by the subsequent re-viewing and re-sharing of this content.

Our third measure (ICS F3), on crisis prevention information, will effectively disrupt user search journeys to minimise the risk of those users encountering illegal suicide content, and minimise the risk of harm should users encounter such content.

# Introduction

9.1 There is evidence that search services can act as a gateway to a wide range of illegal content that is present elsewhere online.[1382] Search functionalities (such as predictive search) that have been designed to optimise search results for relevance and efficiency can also have the unintended consequence of making it easier for users to encounter illegal search content (whether inadvertently or when seeking it out).[1383]

9.2 The Online Safety Act 2023 ('the Act') requires service providers to take steps to minimise the risk of individuals encountering illegal search content, and to effectively mitigate and manage the risks of harm to individuals on their search services.[1384] These duties require providers to take steps, where proportionate, in a number of areas relevant to the design of search services, including (among others):[1385]

- the design of functionalities, algorithms and other features relating to the search engine;
- functionalities allowing users to control the content they encounter in search results;[1386]
- content prioritisation; and
- user support measures.

9.3 The evidence base for measures to tackle illegal harms on search services is more limited than that for user-to-user (U2U) services. As such, we have largely drawn the recommendations in this chapter from existing steps that search service providers take to

[1382] See our Register of Risks ('the Register') chapter titled 'Search services'.

[1383] See our Register chapter titled 'Search services'.

[1384] Section 27(2) of the Act sets out the duty for providers to take or use proportionate measures relating to the design or operation of the service to effectively mitigate and manage the risks of harm to individuals, as identified in the most recent illegal content risk assessment of the service. Section 27(3) of the Act sets out the duty to operate a service using proportionate systems and processes designed to minimise the risk of individuals encountering search content of the following kinds – (a) priority illegal content; (b) other illegal content that the provider knows about.

[1385] Section 27(4)(b), (c), (d) and (e) of the Act.

[1386] We are not recommending measures relevant to functionalities that allow users to control the content they encounter in search results (section 27(4)(c)) in the first iteration of the Codes as we have not seen sufficient evidence that doing so would be effective or proportionate. We will keep this under review as our evidence base develops.

protect their users from harm. Using the research and evidence available, our final measures seek to codify elements of best practice. We expect to build on and refine our approach as we learn more over time.

## Feedback on our approach

9.4 Stakeholders, including civil society organisations and government bodies, expressed broad support for the overall package of measures relating to search settings, functionalities, and user support in the November 2023 Illegal Harms Consultation ('November 2023 Consultation').[1387]

9.5 In our proposals, we distinguished between two kinds of search services:

a) **General search services**, which operate by means of an underlying index of URLs and enable users to search the web by inputting search requests.
b) **Vertical search services**, which enable users to search for specific topics, products or services offered by third-party operators and operate by live querying of selected websites using an Application Programming Interface (API) or equivalent means.

9.6 In the November 2023 Consultation, we explained that there was no clear evidence to suggest that vertical search services play a role in the dissemination of priority illegal content or other illegal content. We have not received clear evidence in consultation responses (or elsewhere) to suggest otherwise in relation to the measures and priority offences discussed in this chapter. We have therefore excluded these services from the scope of the measures in this chapter.[1388] Mid Size Platform Group explicitly expressed support for this approach in their November 2023 Consultation response due to the limited functionalities of vertical search services and the lack of available evidence for harm on those services.[1389]

9.7 Within general search services, we distinguish between services that rely on their own indexing and those which acquire their index or search results from a third-party general search service that does its own indexing (known as downstream general search services). We provide a full description of these types of service in the chapter titled 'Overview of regulated services', and the approach we have taken towards downstream general search services in the Codes, in the chapter titled 'Our approach to developing Codes measures.'

9.8 In the remainder of this chapter, we explain our decisions to include three measures in the Illegal Content Codes of Practice for search services relating to search settings, functionalities and user support on:

[1387] Betting and Gaming Council response to the November 2023 Illegal Harms Consultation, p.14; Canadian Centre for Child Protection (C3P) response to November 2023 Illegal Harms Consultation, p.29; Centro de Estudios en Libertad de Expresion (CELE) response to November 2023 Illegal Harms Consultation, p.14; [✂]; Duran Benjamin O'Dwyer response to November 2023 Illegal Harms Consultation, p.11; Local Government Association response to November 2023 Illegal Harms Consultation, p.16; Mencap response to November 2023 Illegal Harms Consultation, p.16; [✂]; National Trading Standards eCrime team response to November 2023 Illegal Harms Consultation, p.15; Nexus response to November 2023 Illegal Harms Consultation, p.20; Segregated Payments Ltd response to November 2023 Illegal Harms Consultation, p.15; The Cyber Helpline response to November 2023 Illegal Harms Consultation, p.21; Welsh Government response to November 2023 Illegal Harms Consultation, p.5.

[1388] See our Register chapter titled 'Search services'.

[1389] Mid Size Platform Group response to November 2023 Illegal Harms Consultation, p.3.

a) the reporting and removal of predictive search suggestions;
b) the provision of CSAM warnings;
c) the provision of crisis prevention information.

9.9 We set out what we proposed in our November 2023 Consultation, the stakeholder feedback on the proposed measures, our decisions and our reasoning.

## Measure on reporting and removal of predictive search suggestions

9.10 In our November 2023 Consultation, we proposed that providers of large general search services with a predictive search functionality should ensure that users have a means to easily report predictive search suggestions which they believe may direct users towards illegal content.[1390] When a report is received, we proposed the provider should consider whether the wording of the suggestion presents a “clear and logical” risk of users encountering search content that is priority illegal content. If the provider identifies such a risk, it should not recommend the reported predictive search suggestion to any user.

9.11 The aim of this proposed measure was to reduce the risk of predictive search functionalities suggesting search terms which could lead users to encounter harmful and potentially illegal content.

### Summary of stakeholder feedback[1391]

9.12 Two civil society stakeholders expressed their support for this measure.[1392] The 5Rights Foundation agreed that users should be able to access prominently displayed reporting mechanisms given the “known risk” of predictive search.[1393] The National Trading Standards eCrime team noted the potential benefits of the measure for mitigating users accessing fraudulent content.[1394]

9.13 We identified several themes in responses to the November 2023 Consultation, and the May 2024 Consultation on Protecting Children from Harms Online (‘May 2024 Consultation’), relating to the reporting and removal of predictive search suggestions. These included:[1395]

[1390] In the November 2023 Consultation, we referred to this as measure 7A. In our final decision, it is measure ICS F1.

[1391] Note this list is not exhaustive, and further responses can be found in Annex 1.

[1392] In addition, four stakeholders expressed support for the measure’s counterpart in the May 2024 Consultation on Protecting Children from Harms Online (measure SD1): Centre for Excellence for Children’s Care and Protection (CELCIS) response to May 2024 Consultation on Protecting Children from Harms Online, p.18; Jamie Dean response to May 2024 Consultation on Protecting Children from Harms Online, p.20; NSPCC response to May 2024 Consultation on Protecting Children from Harms Online, p.70; Scottish Government response to May 2024 Consultation on Protecting Children from Harms Online, p.19.

[1393] 5Rights Foundation response to November 2023 Illegal Harms Consultation, p.33.

[1394] National Trading Standards eCrime team response to November 2023 Consultation, p.15.

[1395] In May 2024 we proposed measure SD1, which set out that providers should offer users a means to easily report predictive search suggestions which they consider increase the risk of user exposure to primary priority content or priority content harmful to children. See PCS E2 in Annex 8: Protection of Children Code of Practice for search services in the May 2024 consultation. [accessed 18 November 2024].

- the feasibility of determining whether a predictive search suggestion presents a clear and logical risk of directing users towards illegal content in practice;[1396]
- how providers display the options to report predictive search suggestions;[1397]
- the importance of placing responsibility for safety onto service providers, not users;[1398] and
- the services that the measure applies to.[1399]

9.14 We outline these stakeholder concerns in the following sections, and address additional stakeholder responses in Annex 1.

### Determining whether a predictive search suggestion presents a 'clear and logical' risk

9.15 Some stakeholders expressed concerns about the process of determining whether a predictive search suggestion presents a risk of directing users towards illegal content.[1400] Google expressed concerns that the "clear and logical risk" threshold for actioning reported predictive search suggestions was too low, and argued that the nature of search suggestions would make it challenging (if not impossible) to rule out the possibility of users encountering illegal content when clicking on them.[1401] [✂].[1402] We address this concern in paragraph 9.23 in the section entitled 'How this measure works'.

### Display of reporting options

9.16 We also received some relevant feedback on the measure on the reporting and removal of predictive search suggestions that we proposed in the May 2024 Consultation. The Canadian Centre for Child Protection (C3P) made suggestions about the display of the recommended reporting tools. This included that we should clarify where and when service providers should make the reporting tool available to users, for example, making it available without the user having to carry out a search. It also suggested that we should recommend a standardised approach to how providers display reporting tools, such as setting a minimum font size and colour to ensure the options are easily accessible.[1403] We address this concern in paragraph 9.35 in the section entitled 'Benefits and effectiveness'.

### Responsibility for safety

9.17 While the 5Rights Foundation agreed with the measure in principle, it argued that it should be considered "complementary" to other safety by design measures that shift the responsibility for safety onto service providers (rather than onto users).[1404] We address this concern in paragraph 9.36 in the section entitled 'Benefits and effectiveness'.

[1396] Google response to November 2023 Illegal Harms Consultation, p.70; [✂].
[1397] Canadian Centre for Child Protection (C3P) response to May 2024 Consultation on Protecting Children from Harms Online, pp.32-33.
[1398] 5Rights Foundation response to response to November 2023 Consultation, p.33.
[1399] C3P response to May 2024 Consultation, p.32.
[1400] Google response to November 2023 Consultation, p.70; [✂].
[1401] Google response to November 2023 Consultation, p.70.
[1402] [✂].
[1403] C3P response to May 2024 Consultation, pp.32-33.
[1404] 5Rights Foundation response to response to November 2023 Consultation, p.33.

### Scope of the measure

9.18 In its response, C3P also called for the predictive search measure in the May 2024 Consultation to extend to all general search services, regardless of size, to ensure that all services meet their basic safety obligations.[1405] We address this concern in paragraph 9.60 in the section entitled 'Who this measure applies to'.

## Our decision

9.19 We have decided to broadly confirm the measure we proposed in our November 2023 Consultation. We have made a minor clarifying change in response to the feedback we have received:

- In our November 2023 proposal, we proposed that providers should take steps to ensure predictive search suggestions that present a "clear and logical" risk of directing users to illegal content are no longer recommended to users. Stakeholders interpreted this threshold to be lower than we originally intended. We have therefore reframed this threshold as "clear and material".

9.20 Our measure now says:

a) Providers of large general search services that use a predictive search functionality, should offer users a means to easily report predictive search suggestions which they consider direct users towards priority illegal content. Where a report is received, the provider should:

   i) consider whether the wording of a reported predictive search suggestion presents a clear and material risk of users encountering illegal content; and

   ii) if a risk is identified, take appropriate steps to ensure that the reported predictive search suggestion is not recommended to any user.

9.21 The full draft of the measure is included in the Illegal Content Codes of Practice for search services on terrorism, CSEA and other duties and is referred to as measure ICS F1.

## Our reasoning

### How this measure works

9.22 Under this measure, providers of search services should offer users an easy way to report predictive search suggestions which they think may direct users towards priority illegal content. The provider should consider whether the wording of each reported predictive search suggestion presents a clear and material risk of users encountering illegal content. They should then take appropriate steps to ensure that the reported predictive search suggestion is not recommended to any further users. This should ensure that they no longer present users with harmful search suggestions. In turn, this should minimise the risk of users encountering illegal content in search results by clicking on a predictive search suggestion.

9.23 In developing this measure, we intended the original "clear and logical" threshold that we consulted on to be higher than some stakeholders have interpreted.[1406] We do not expect providers to remove every reported predictive search suggestion that presents any risk of

[1405] C3P response to May 2024 Consultation, p.32.

[1406] Google response to November 2023 Consultation, p.70.

users encountering illegal content should they click on it. For example, there may be instances where a predictive search suggestion presents a clear risk of encountering illegal content, but that risk is unlikely to materialise in practice. We have therefore amended our measure in response to this feedback to make clear that we expect service providers to determine whether the risk of "encountering illegal content" via the predictive search suggestion is "clear" and "material" based on the wording of the suggestion. We removed the additional qualifier "logical" for clarity as we considered that it served the same function as the word "clear". In this context:

- "Encountering illegal content" refers both to illegal content that users come across via the featured snippets available in the search results and to content they might come across within one click through the blue hyperlinks in the search results. We do not expect providers to consider the risk of encountering illegal content on user journeys more than one click away from the search results.
- "Clear" refers to there being an obvious risk of encountering illegal content based on the wording of the predictive search suggestion. This determination should be reasonable and based on good judgement.
- "Material" refers to it being likely that the predictive search suggestion will lead to illegal content in practice. Where a suggested query is unlikely to return illegal content, we would not expect providers to take steps to ensure that the predictive search suggestion is no longer recommended.

9.24 We offer flexibility for providers both in how they determine whether a predictive search suggestion presents a clear and material risk of directing users to illegal content and in how they ensure it is not then recommended to any users where appropriate. We are not expecting providers to conduct analysis of the search results returned by every reported predictive search suggestion to assess whether and how much illegal content is returned (though they may choose to do so if they consider it appropriate).

## Benefits and effectiveness

### Benefits

9.25 In this section, we describe the benefits we consider this measure will bring to address risks relating to predictive search functionalities. Predictive search functionalities are algorithmic features embedded in the search bar of a search service.

9.26 When a user begins to input a search request, the algorithm predicts the rest of the request and suggests possible related search terms to help users make more relevant searches. Predictions are based on many factors including a user's past queries, other user queries, locations, and trends.[1407] Several search services use these functionalities, with well-known examples including Google Search's autocomplete functionality and Microsoft Bing's autosuggest tool.

9.27 Predictive search can increase the risk of individuals receiving search suggestions that direct them to illegal content. This is because a predictive search suggestion might prompt a user to search for illegal content that they might otherwise not have searched for had the query not been suggested.

[1407] Google Search Help. How Google autocomplete predictions work. [accessed 18 November 2024]; Microsoft Support. How Bing delivers search results. [accessed 18 November 2024].

9.28 Our Register of Risks ('Register') chapter titled 'Search services' presents evidence of how predictive search can increase the risk of users encountering content relating to child sexual abuse material (CSAM), fraud, hate, and instructions for self-harm and suicide (which, depending on the context, may amount to illegal content). We consider it reasonable to assume that predictive search could also make it easier for users to encounter search content that is illegal content in other priority offence areas.

9.29 This measure seeks to address the risk of encountering priority illegal content via predictive search. It is likely to be most beneficial for users who are not actively searching for illegal content but who may inadvertently (or out of curiosity), when prompted, click on a predictive search suggestion that leads to it. This is because the measure will reduce the risk of users being presented with search suggestions that might lead them to encounter illegal content.

9.30 There may also be an incidental benefit to the wellbeing of users who would be distressed by a suggestion that clearly directs users to encounter illegal content, but who would not take further action to search for illegal content when prompted.[1408]

9.31 Users actively searching for illegal content are unlikely to receive any significant benefit from this measure, as they can still type in specific search requests to locate the results they want. That said, the measure may provide a small benefit in reducing the ease of accessibility of illegal content to this user group by adding further friction into the user journey.

### Effectiveness

9.32 The effectiveness of the measure in addressing the risk of encountering illegal content occurs in two stages. These are i) enabling user reporting, and ii) ensuring that predictive search suggestions that are likely to direct users to illegal content are no longer recommended to them.

9.33 As we explained in our November 2023 Consultation, current industry practice indicates that this measure is a technically feasible way for the providers of general search services to minimise the risk of individuals encountering priority illegal content via the predictive search functionality, supporting their duties under section 27 of the Act.[1409]

9.34 The first way in which our measure addresses the risk of encountering illegal content is via user reporting. Our wider research into reporting and complaints processes suggests that

[1408] There is evidence of this occurring in relation to CSAM. Source: Constine, J., 2019. Microsoft Bing not only shows child sexual abuse, it suggests it, Tech Crunch, 10 January 2019. [accessed 18 November 2024].

[1409] Both Google and Microsoft already enable user complaints related to predictive search suggestions and take steps to detect and manage search suggestions that are harmful or that violate their policies on Google Search and Microsoft Bing, respectively. Google says it removes autocomplete suggestions that violate its general or specific autocomplete policies (where these are not caught by its automated systems designed to prevent the suggestion of harmful queries), including where predictions contain dangerous, harassing, hateful, or terrorist content. Source: Google Search Help, How Google autocomplete predictions work. [accessed 18 November 2024]. Microsoft aims to prevent users being inadvertently exposed to "potentially harmful, offensive, or misleading content" via search suggestions. It does so through a mixture of proactive and reactive interventions, and also allows users to turn search suggestions on or off. Source: Microsoft Support, How Bing delivers search results. [accessed 18 November 2024]. Some smaller search services like Yahoo also have predictive search functionalities and reporting systems in place for users to report inappropriate predictions. Source: Yahoo! Help, About Yahoo Search Predictions. [accessed 18 November 2024].

users are less likely to engage with such processes or make reports where the process is not easy to find or use.[1410] Providing users with a means to easily report predictive search queries under this measure should improve current practice by effectively reducing barriers to reporting.[1411] Doing so will in turn raise providers' awareness of problematic search suggestions that might otherwise remain undetected.

9.35 We agree with C3P that tools for reporting should be easily accessible.[1412] A reporting tool that is only available after a user has conducted a search would not, in our view, be easily accessible if the positioning of the tool required users to first click through a reported search suggestions and scroll through search results that may contain illegal content before they can report. We would therefore expect providers to consider whether the tool to report a predictive search suggestion is accessible to the user without risking the user encountering potentially illegal search results, to sufficiently secure this provision. We have not gone further to recommend a prescriptive approach to displaying reporting options at this stage. This is because we consider it appropriate to give providers discretion to design their reporting functions (for example, to align with their branding) so long as they allow for users to easily report predictive search suggestions. We encourage providers to consider any factors, such as the location of any signposting to the reporting function, the point at which it appears in the search journey, and the size and colour of any text, that will be helpful in achieving this outcome.

9.36 We understand that some providers of general search services already use automated systems to identify and prevent potentially violative predictions (in addition to user reporting).[1413] We have decided not to recommend automated technical approaches under this measure, though we note that doing so might address the 5Rights Foundation's concern that measures focusing on user controls and tools should be considered complementary to other safety design measures that make providers responsible for the safety of their services.[1414] Due to limited information on the technical operations and underlying policies governing these approaches, it is not clear to us that such a recommendation would be effective or proportionate. We will keep the evidence available under review and may consider a more proactive approach in future if warranted.

9.37 Improved reporting processes under this measure may help service providers to strengthen their existing proactive moderation systems using the outcomes of predictive search reports. For example, Google explained that it already aggregates and collects violative predictive search suggestions reported by users to help improve the algorithms that underpin its proactive moderation of predictive search suggestions over time.[1415] We

[1410] See chapter 6 of this Volume: 'Reporting and complaints'.

[1411] For example, Google Search gives users the option to "report inappropriate predictions" on the search bar. This option is written in grey italics at the bottom of the suggestion box, in a font size smaller than the text it surrounds. Microsoft Bing offers an option to provide feedback on its 'suggest' feature at the bottom of the search results page, or through clicking on the "settings" option at the side of the homepage and scrolling half-way down to click on a "feedback" option. Source: Ofcom desk research, conducted 18 November 2024.

[1412] C3P response to May 2024 consultation, pp.32-33.

[1413] Google Search Help, How Google autocomplete predictions work. [accessed 18 November 2024]; Microsoft Support, How Bing delivers search results. [accessed 18 November 2024].

[1414] 5Rights Foundation response to response to November 2023 Consultation, p.33.

[1415] Ofcom/Google meeting, 24 July 2024, subsequently confirmed by Google by email on 9 August 2024.

discuss how providers' existing automated systems might help to limit the costs of implementing this measure in paragraph 9.45 in the section titled 'Costs and risks'.

9.38 The second way that the measure addresses the risk of encountering illegal content is through ensuring that predictive search suggestions that present a risk of directing users to illegal content are no longer recommended to them. This reduces the future likelihood of other users seeing that suggestion, being prompted to run the search, and potentially encountering illegal content as a result.

9.39 A 2019 report by the Antisemitism Policy Trust and Community Security Trust found that Google's removal of a specific antisemitic predictive search suggestion resulted in 10% (one in ten) fewer search requests related to that suggestion in the 12 months following its removal compared to the 12 months prior.[1416] This indicates that the removal of suggestions deemed to present an illegal content risk could materially reduce the likelihood of users encountering illegal content in search results.

9.40 In summary, this measure provides meaningful benefits by reducing barriers to reporting predictive search suggestions and reducing the likelihood of users encountering priority illegal content in the search results. This will help providers to meet their duties to minimise the risk of users encountering illegal search content under section 27(3) of the Act. In turn, this will help providers to meet their duties to mitigate and manage the risks of harm to individuals under section 27(2).

## Costs and risks

### Costs

9.41 The Act requires general search service providers to implement complaints and reporting systems (see chapter 6 of this Volume: 'Reporting and complaints') to cover a wide range of topics. Any additional costs related to this measure will be the incremental costs to ensure users can also easily report predictive search suggestions and providers can take appropriate action via the reporting mechanisms.[1417] Overall, we expect the extension of these systems to be relatively straightforward.[1418]

9.42 We expect the implementation of this measure to take approximately 20 to 40 days of software engineering time (along with an equal amount of time input from professional occupation staff). We estimate one-off implementation costs to be £10,000 to £40,000.[1419]

9.43 There will also be some ongoing costs. These will include both the incremental maintenance costs of running the extended reporting system, and the additional moderation costs that providers will incur when responding to reports about predictive search.

[1416] Antisemitism Policy Trust, Community Security Trust (Stephens-Davidowitz, S.). 2019. Hidden Hate: What Google searches tell us about antisemitism today. [accessed 18 November 2024].

[1417] We also note that we proposed a similar measure in the May 2024 Consultation. See Measure PC2 E2 in Annex 8. [accessed 18 November 2024]. In the event that a search service is also subject to this measure, then there are likely to be cost synergies across the two measures.

[1418] For example, where a reporting system already exists, the costs of allowing an end user to report contents of the autosuggest within a search term input box are incremental.

[1419] Based on our labour cost assumptions set out in Annex 5. We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by the Office for National Statistics (ONS). We received some feedback on the general cost assumptions (for example, salary assumptions) that are fed into these costs. We consider that feedback in Annex 5.

9.44 We expect these incremental annual maintenance costs to be equivalent to 25% (one-fourth) of the implementation cost and estimate this to be £2,500 to £10,000 per year.[1420]

9.45 The additional moderation costs to review the reports related to predictive search suggestions will likely vary depending on the size of the service. Large general search services will likely receive a larger number of user reports and therefore require a greater number of moderators. Some search services (for example, Google Search and Microsoft Bing) that have a predictive search functionality already make efforts to moderate this functionality to limit the likelihood of suggesting illegal or harmful content. This includes allowing users to report problems with predictive search, as described in paragraph 9.33. Therefore, we expect that these services will incur negligible or limited additional costs due to this measure unless they intended to remove this functionality.

9.46 Also, providers that use automated moderation of search content (such as Google) may be able to make use of these existing measures to process predictive search reports.[1421] This may further mitigate the increase in moderation costs if the provider already has an automated moderation functionality that is able to handle these types of reports (or can be adapted to do so). Services which do not have predictive search functionality (such as Mojeek) would not be in scope of the measure and so providers of these services would not incur any additional costs from this measure unless they planned to introduce such a function in future.

### Risks

9.47 There is a risk that providers of general search services may remove predictive search functionalities to avoid the consequences of failing to properly moderate them. This could negatively affect the user experience. We expect this risk to be minimal as several services already have this measure in place and already have a reporting process for predictive search suggestions.

9.48 Similarly, there is a risk that providers may over-moderate the predictive search algorithm, which could cause the function to lose functionality. This risk is likely to be minimal as providers have a commercial incentive to ensure that predictive search is informative for users.

9.49 Overall, we consider that the costs and risks of this measure will likely be relatively low. We will continue to gather evidence of the impact of this on smaller services, particularly as the cost of moderating reports is uncertain. We also have limited information on some of the existing smaller search services in the market and their approach to predictive search. We consider that the costs at the upper end of our estimate could potentially be material for smaller services.

[1420] As described in Annex 5, we assume annual maintenance costs are 25% (one-fourth) of the initial costs where we have no more specific information. We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS. We received some feedback on the general cost assumptions (for example, salary assumptions) that are fed into these costs. We consider that feedback in Annex 5.

[1421] Google and YouTube, Information quality & content moderation. [accessed 18 November 2024].

## Rights impact

### Freedom of expression

9.50 As explained in 'Introduction, our duties, and navigating the Statement', and in chapter 14 of this Volume: 'Statutory tests', Article 10 of the European Convention on Human Rights (ECHR) sets out 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 this is a qualified right, we must exercise our duties under the Act in a way that does not restrict this right unless we are satisfied that it is necessary and proportionate to do so.[1422]

9.51 We do not consider that this measure would have any material impact on users' rights to receive information under Article 10. While we acknowledge that the measure will affect the availability of some search suggestions which present a clear and material risk of directing users towards illegal content, the removal of the suggestion would not prevent users from inputting search requests or accessing search results through the service. For the same reasons, we do not consider that actions taken by providers in line with this measure would have any material impact on website operators; the website would remain discoverable via the search service even where a predictive search suggestion that surfaces that URL is removed.

9.52 We consider there to be a very small impact on the freedom of expression rights of the providers of search services, as their right to impart information to users in the form of predictive search suggestions would be restricted. However, the impact would be proportionate to the measure's overall legitimate aim of minimising the risk of users encountering priority illegal content via search services.

9.53 Taking these points into account, we consider the impact of this measure on the right to freedom of expression to be limited and proportionate.[1423]

### Privacy

9.54 As explained in 'Introduction, our duties, and navigating the Statement', and in chapter 14 of this Volume: 'Statutory tests', 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.

9.55 We acknowledge that the processes established to consider and action user reports related to predictive search suggestions could have implications for the right to privacy of the reporting user and their rights under data protection law. The degree of interference with the right to privacy will depend to a degree on the extent to which the nature of the affected content gives rise to a legitimate expectation of privacy.

9.56 This measure gives service providers flexibility as to precisely how they design the reporting tool and handle reports. Depending on how they do so, the submission of a report may involve the provision of additional information to contextualise the reason for the report, and the handling of a report by the provider may involve the processing of personal data relating to the reporting user. To the extent that this information is provided freely by the

[1422] A qualified right is a right that can be restricted in certain circumstances to balance the rights of the individual with the needs of another, or of the wider community.

[1423] Above and beyond the requirements of the Act.

reporting user to the provider, we consider that any interference with the right to privacy involved in the implementation of this measure would be limited and proportionate to the legitimate aim of the measure.

### Data protection

9.57 We recognise that the processes established to implement this measure, as outlined above, might generate new personal data or involve processing existing data for new purposes (if the service provider considers it appropriate to retain information about the reporting user or the reports themselves – for example, for prioritisation or training purposes). However, this measure does not suggest or require that providers retain any reporting user's personal data. Where the reporting mechanisms put in place to implement this measure involve personal data processing, providers must comply with relevant data protection legislation. This includes applying appropriate safeguards to protect the rights of both children (who may require special consideration) and adults who may submit reports regarding predictive search suggestions. Providers should refer to relevant guidance from the Information Commissioner's Office (ICO).[1424]

9.58 We consider the impact of this measure on the privacy rights of users is limited where providers comply with relevant data protection laws. As compliance with this measure will aid in satisfying the provider's duties under the Act, if there is an interference it is proportionate to the safety benefits to users.

## Who this measure applies to

9.59 In our November 2023 Consultation, we proposed recommending this measure to providers of large general search services that use predictive search functionalities. We considered the measure to be proportionate for such services as the costs are likely to be relatively small compared to the potential benefits in reducing the risk of harm.

9.60 As discussed in paragraph 9.18, we received feedback from C3P suggesting that this measure should extend beyond large search services to apply to all search services, regardless of size.[1425] However, we do not consider there to be sufficient evidence to support extending this measure to smaller general search services at this stage:

- The benefits of applying this measure to a service with limited reach are likely to be relatively small. As this measure is expected to benefit users who are not looking for harmful content, there is limited risk of displacing users from large services to small services. This reduces the need to apply the measure to smaller services.
- Although the costs are likely to be relatively limited, we expect that, in some scenarios, the costs could still be material for a smaller general search service.

9.61 We therefore consider that this measure is appropriate for large general search services that use predictive search functionalities.

# Conclusion

9.62 In light of this analysis, we conclude that applying this measure to large general search services will produce meaningful benefits by reducing barriers to reporting predictive

[1424] ICO, UK GDPR guidance and resources; and Online safety and data protection. [accessed 18 November 2024].

[1425] C3P response to May 2024 Consultation, p.32.

search suggestions and reducing the likelihood of users encountering priority illegal content in the search results. At the same time, it should produce limited costs. We have therefore decided that this measure should apply to providers of large general search services that use predictive search functionalities.

9.63 This measure is included in our Illegal Content Codes of Practice for search services on terrorism, CSEA, and other duties. It is referred to within these Codes as 'ICS F1'.

# Measure on provision of CSAM content warnings

9.64 In the November 2023 Consultation, we proposed that the providers of large general search services should employ means to detect search requests where the wording clearly indicates that the user may be seeking to encounter CSAM or search requests that use terms that explicitly relate to CSAM. Service providers should provide warning messages in response to these search requests.[1426]

9.65 The aim of the proposed measure was to deter users from seeking to encounter CSAM, thereby reducing the risk of users encountering illegal search content and the broader harm that might result from this.

## Summary of stakeholder feedback[1427]

9.66 In addition to those stakeholders who expressed broader support for the full package of search settings, functionalities, and user support measures outlined in paragraph 9.4, two civil society stakeholders expressed support specifically for this measure.[1428]

9.67 Our analysis of the responses identified several areas where stakeholders argued this measure could go further. These included (but were not limited to):

- the practices around establishing and maintaining a list of CSAM terms;[1429]
- the accessibility and effectiveness of the warning message and its content;[1430]
- our assessment of the impact of the measure on privacy;[1431] and
- applying the measure to all search services.[1432]

9.68 We outline these stakeholder concerns in more detail in the following section.

[1426] In the November 2023 Consultation, we referred to this measure as Measure 7B.

[1427] Note this list is not exhaustive, and further responses can be found in Annex 1.

[1428] The NSPCC noted the overall importance of signposting measures, while WeProtect Global Alliance described warnings and information as a "key part of a preventative approach to reducing child sexual exploitation and abuse online." NSPCC response to November 2023 Illegal Harms Consultation, p.45; WeProtect Global Alliance response to November 2023 Illegal Harms Consultation, p.26.

[1429] 5Rights Foundation response to November 2023 Consultation, p.33, BILETA response to November 2023 Illegal Harms Consultation, p.19, NSPCC response to November 2023 Consultation, p.45, Protection Group international response to November 2023 Illegal Harms Consultation, p.13.

[1430] [✂].

[1431] ICO response to November 2023 Illegal Harms Consultation, p.21.

[1432] C3P response to November 2023 Consultation, p.30; Lucy Faithfull Foundation response to November 2023 Consultation, p.8; WeProtect Global Alliance response to November 2023 Consultation, p.26.

## Establishing and maintaining a list of CSAM terms

9.69 Some civil society organisations suggested ways in which the measure could improve existing practices around establishing and maintaining a list of CSAM terms. This included the suggestion that service providers might achieve greater consistency by sharing new terms amongst themselves or by nominating a central provider to co-ordinate the list.[1433] The British and Irish Law and Technology Association (BILETA) also noted the potential for inconsistencies between service providers' policies relating to their CSAM warning messages more generally.[1434]

9.70 We address these points in paragraphs 9.86 to 9.89 under the section entitled 'How this measure works'.

## Accessibility and effectiveness of the warning message

9.71 One stakeholder suggested that there should be different links to appropriate resources for adults and children in recognition of the increasing number of child sexual offences committed by children.[1435] We address this concern in paragraph 9.92 under the section entitled 'How this measure works'.

9.72 The Lucy Faithfull Foundation recommended that service providers should draw on experts in deterrence messaging, such as civil society organisations, to support the improved effectiveness of warnings. It noted that how a warning message appears and what it says are important factors relevant to effectiveness.[1436]

9.73 The Lucy Faithfull Foundation provided two anecdotal examples of how it considers expertise could strengthen the effectiveness of a warning message. In the first, it described how one search service had removed signposting to the 'Stop it Now!' website, cutting off referrals to that support website. It provided some quantitative data in relation to the 'Stop It Now!' website to demonstrate that, after working with the provider of that search service to change the message and reintroduce the signposting, traffic to that website increased at a higher rate than before. In its second example, the Lucy Faithfull Foundation pointed out some problems with another provider's warning message, noting that it had not recently worked with that provider on its message. In particular, it noted that the warning message appeared at the bottom of the search results and therefore did not stand out.[1437] It argued that these anecdotal examples are evidence of the need for civil society organisations, who are experts in deterrence messaging, to play a role in the measure. We note that, since the November 2023 Consultation, the Lucy Faithfull Foundation has worked with this service provider to improve their warning, which now appears at the top of the search results page.[1438]

9.74 We address this concern in paragraphs 9.99 to 9.100 under the section entitled 'How this measure works'.

[1433] NSPCC response to November 2023 Consultation, p.45; Protection Group International Inc. response to November 2023 Consultation, p.13. 5Rights Foundation also called for Ofcom to consider the NSPCC's response to this measure. Source: 5Rights Foundation response to November 2023 Consultation, p.33.

[1434] BILETA response to November 2023 Consultation, p.19.

[1435] [✂].

[1436] Lucy Faithfull Foundation response to November 2023 Consultation, pp.6-8.

[1437] Lucy Faithfull Foundation response to November 2023 Consultation, pp. 7-8.

[1438] Lucy Faithfull Foundation email dated 21 November 2023.

9.75 Though BILETA agreed with the provision of CSAM warnings in principle, it expressed concerns about the effectiveness of such warnings in deterring users who are actively seeking out CSAM content.[1439] We address this concern in paragraph 9.108 under the section entitled 'Benefits and effectiveness'.

### Feedback on our rights assessment

9.76 The ICO recommended that we review our assessment of the impact of the measure on the right to privacy to ensure that it fully captures the relevant data protection considerations.[1440] It argued that the measure could lead to providers processing personal data to deliver warnings to individual identifiable users (depending on how providers implement CSAM warnings).[1441] We address this concern in paragraph 9.136 under the section entitled 'Rights impact'.

### Scope of the measure

9.77 Three stakeholders proposed that this measure should extend to search functions on services beyond only large general search services.[1442] The Lucy Faithfull Foundation provided evidence of providers deploying similar initiatives within U2U services in support of this view.[1443] We address this concern in paragraph 9.139 under the section entitled 'Who this measure applies to'.

## Our decision

9.78 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation. We have made a small number of changes in response to the feedback received:

- We have amended this measure to recommend that the CSAM warning information should be "comprehensible and suitable in tone and content for as many UK users as possible, including children".
- We have amended the measure to recommend that service providers develop their warning messages with input from a person with expertise in deterring CSEA offences.

9.79 Our measure now says:

a) Providers of large general search services should employ means to detect and provide warnings in response to search requests where both (a) the wording of the search requests indicates that a user may be seeking to encounter CSAM; and (b) terms or combinations of letters and symbols that explicitly relate to CSAM, are used in the search request.

i) The warning should provide information regarding the illegality of CSAM that is comprehensible and suitable in tone and content for as many users as possible, including children, and link(s) to resources designed to help users refrain from

[1439] BILETA response to November 2023 Consultation, p.18.
[1440] ICO response to November 2023 Consultation, p.21.
[1441] ICO response to November 2023 Consultation, p.21.
[1442] C3P response to November 2023 Consultation, p.29; Lucy Faithfull Foundation response to November 2023 Consultation, p.8; WeProtect Global Alliance response to November 2023 Consultation, p.26.
[1443] Lucy Faithfull Foundation response to November 2023 Consultation, p.8.

committing CSAM offences. It should be developed with input from a person with expertise in deterring CSEA offences.

9.80 This measure is included in our Illegal Content Codes of Practice for search services and is referred to as measure ICS F2.

# Our reasoning

## How this measure works

9.81 Under this measure, providers of general search services should display content warnings and support resources in response to user searches for CSAM. They should do so by:

a) establishing and maintaining a list of CSAM terms;
b) detecting relevant terms; and
c) displaying a warning message and links to support resources when a user inputs a search query associated with CSAM.

### CSAM search terms

9.82 This measure recommends that providers deploy a warning message if they detect relevant terms within a user's search query. Perpetrators use many different search terms of varying specificity to locate CSAM, and it may not be appropriate to display a warning in every case. Broadly, we consider the terms used by perpetrators to search for CSAM to fall into three categories:

a) Obvious CSAM-specific terms that would be understandable to the lay person and clearly indicate that a user is seeking to encounter CSAM.
b) CSAM-specific terms that are used among offenders to evade detection, which would not be understood by a lay person to relate to CSAM (for example, code words or terms made up of combinations of letters and symbols).
c) Seemingly innocuous terms that are known among perpetrators to generate CSAM results but are not CSAM-specific.

9.83 We recommend that service providers establish and maintain a list of CSAM terms which fall under categories (a) and (b) only. This is because terms in these categories comprise CSAM-specific terms (including words or combinations of letters and symbols) and would be unlikely to be used to generate non-CSAM results. Their use clearly indicates that a user is seeking to access CSAM. However, we do not consider it appropriate for the list to include terms that fall under category (c). Doing so could have the severe unintended consequences of informing the user that the term they have entered is CSAM-related, inadvertently making them aware of how they might search for CSAM if they wished to do so. While we acknowledge this approach will narrow the application of this measure, we consider it to be effective at targeting perpetrators using some coded terms, while avoiding this potentially severe risk.

9.84 Service providers may choose to develop their own list of terms for which they consider it appropriate to provide a warning. Alternatively, they may wish to source a list of terms from an expert third-party person or organisation or use a combination of these two approaches. The current practice outlined in footnote 1461 suggests that any of these approaches is feasible.

9.85 Regardless of the approach taken, providers should consider several principles in developing or sourcing a list of CSAM search terms:

- The list should be developed by or sourced from a person (or organisation) with expertise in terms commonly used by offenders to search for CSAM online.
- The list should be regularly updated to add newly discovered terms and to remove terms that are no longer relevant.
- Only search terms meeting the description in paragraph 9.82 (a) and (b) should be used for the purposes of this measure, while the terms described at paragraph 9.82 (c) should be omitted.[1444]
- The list should be secured against unauthorised access, interference, or exploitation by potential perpetrators who may seek to obtain the list to discover and share terms that can be used to search for CSAM. Methods of securing the list could include procedural, physical, human, and technical controls.

9.86 We acknowledge the potential benefits of service providers sharing collective insights on evolving terms to provide more comprehensive coverage. As highlighted by Protection Group International and C3P, new CSAM terms may emerge through perpetrators seeking to avoid detection, and sharing terms could ensure all providers issue warnings based on up-to-date intelligence.[1445]

9.87 However, the evidence provided by stakeholders indicates that the benefits of shared lists would be greatest in relation to the seemingly innocuous terms in category (c) at paragraph 9.82, which we do not recommend using as part of this measure. While we expect there would be some benefits relating to terms in category (b), the benefits relating to category (a) would be very limited as we do not expect to see frequent language shifts in this area. Furthermore, the sensitive nature of the terms means that establishing and maintaining a process for sharing them may incur additional costs due to the need to establish sufficient safeguards. At this stage, we have not seen sufficient evidence to reach a clear conclusion on whether the potential benefits of recommending service providers should share lists as part of this measure would be proportionate to the likely costs and risks.

9.88 This measure allows service providers the flexibility to use a third-party list of CSAM terms if they consider it appropriate. While we acknowledge that there are expert third-party organisations who can usefully advise service providers in this area, there are several reasons why we do not think it is appropriate to prescribe that the list of CSAM terms is co-ordinated by a centralised list provider as part of the measure:

- We do not have access to information on every active third-party list provider that would allow us to recommend one organisation over others who may be capable of performing the same function.[1446]
- We want to ensure that the measure is future-proof. Recommending a specific centralised list provider could undermine this aim as the availability of third-party list

[1444] We understand that many third-party keyword lists for CSAM are developed for the purposes of content moderation, and as such, may contain terms that fall within the category identified in paragraph 9.82(c), which we do not recommend that providers include as part of this measure.

[1445] C3P response to November 2023 Consultation, p.29; Protection Group International response to November 2023 Consultation, p.13.

[1446] In addition, see footnote 1473 which explains why some third-party lists may not be suitable for this purpose.

providers is subject to change over time. This approach would require extensive oversight to ensure quality and effectiveness.[1447]

- Recommending a centralised third-party list provider would limit the ability of service providers to tailor lists to their services.

9.89 It is not clear to us that the potential benefits of a centralised list outweigh the practical challenges and potential costs. That said, we will keep this under review and may reconsider our position in future, based on further evidence and analysis.

### CSAM warning message

9.90 Service providers should ensure that the CSAM warning message:

- provides information on the illegality of CSAM that is suitable and comprehensible in tone for as many UK users as possible, including children;
- provides links to resources designed to provide support and information to help users who are purposefully seeking out such material but wish to stop doing so;
- is developed with input from person(s) with expertise in deterring CSEA offences; and
- is prominently displayed so that it is the first piece of information users encounter in search results (this could be done in different ways, such as by displaying a pop-up or banner).[1448]

9.91 Codifying these criteria ensures more consistent policies between difference search services in this area. We expect this should help address BILETA's concern (noted in paragraph 9.69).[1449]

9.92 Given the evidence suggesting that child-on-child sexual offences are increasing, we agree there are merits to the suggestion from one stakeholder that the warning message should signpost users to age-appropriate links.[1450] A 2022 report analysing all police-recorded CSEA crime suggested that child-on-child abuse accounted for just over half of all CSEA offences that year.[1451] This indicates that some users searching for CSAM via search services may be children.

9.93 As these child users would receive warnings in accordance with this measure, it may be useful for providers to direct children seeking out CSAM to age-appropriate support materials that can help them stop this behaviour. However, we do not consider it proportionate or feasible to recommend this in the measure for two reasons.

9.94 Firstly, on further exploration, the current availability of resources designed to help children refrain from committing CSEA offences appears to be limited. We are aware of three

[1447] Any centralised list or sharing of terms would need sufficient safeguards in place. If a centralised list was to leak, for example, this would make it much easier for users to evade detection where all services are using the same list.

[1448] A pop-up refers to a window or dialog box that appears in the foreground of the window that a user is accessing in response to user action.

[1449] BILETA response to November 2023 Consultation, p.19.

[1450] [✂].

[1451] National Police Chiefs' Council, 2023. National Analysis of Police-Recorded Child Sexual Abuse & Exploitation (CSAE) Crimes Report, p.8. [accessed 18 November 2024]. The report covers a broader spectrum of offences than our measure, including CSAM offences such as the taking, making, and sharing indecent images of children (which was found to be the third most common child-on-child offence).

resources, two of which appear to be targeted at a US audience, and one relatively new resource (shorespace.org.uk) which has not yet been the subject of any signposting interventions targeted at children.[1452]

9.95 Secondly, we do not consider it would be feasible or proportionate to require service providers to display different messages tailored to adults and children. While general search services may use other tools such as user profiling technologies, we understand that the primary means of establishing the age of users is through self-declaration, which the Act explicitly states is not a form of age assurance.[1453] This may be unlikely to change in the immediate future as we proposed not to recommend highly effective age assurance for search services under the Codes at this stage. This proposal is subject to Ofcom's consideration of responses to the May 2024 Consultation.[1454]

9.96 Without the means for tailored messages, a provider would need to present child-specific information alongside any other links to resources for users of all ages, including adults. This could make the warning message more cumbersome and reduce its effectiveness.

9.97 We instead consider that there is an opportunity for service providers to improve the accessibility of the warning message content itself to ensure it reaches as many users as possible. We have therefore amended this measure to recommend that the CSAM warning information should be comprehensible and suitable in tone and content for as many UK users as possible, including children.

9.98 We are not prescriptive about how the warning message should be provided, and therefore, at this stage we do not think it is appropriate to specify particular ways in which it should be accessible to disabled people. However, providers should consider their obligations under other relevant legislation (for example, the Equality Act 2010) and (where relevant) appropriate guidance.[1455]

9.99 We consider it reasonable to expect that a warning message developed in collaboration with experts (as suggested by the Lucy Faithfull Foundation) would be more effective at deterring searches for CSAM and providing appropriate support than a message developed without this collaboration, all else being equal.[1456] There are constraints on the evidence available in this area given the difficulties in quantifying the effect a particular message might have on potential perpetrators. However, in response to the November 2023 Consultation, the Lucy Faithfull Foundation provided anecdotal evidence of how the involvement of expertise could strengthen the effectiveness of warning messages.[1457]

9.100 We have decided to amend the measure to recommend that service providers develop their warning messages with input from a person with expertise in deterring CSEA offences. We would expect providers to seek input on the drafting and positioning of the message itself, as well as on the appropriateness of the resources it signposts to. While we recognise the expertise within civil society organisations, we have chosen not to specify whether service providers should consult with internal or external experts. We have taken this

[1452] Ofcom/ Lucy Faithfull Foundation meeting, 27 June 2024, subsequently confirmed by Lucy Faithfull Foundation by email on 6 August 2024.

[1453] Section 230(4) of the Act.

[1454] See Volume 5 of the May 2024 Consultation, paragraph 15.8.

[1455] For example, see the Web Accessibility Initiative. WCAG 2 Overview. [accessed 18 November 2024].

[1456] Lucy Faithfull Foundation response to November 2023 Consultation, p.6.

[1457] Lucy Faithfull Foundation response to November 2023 Consultation, pp.7-8.

approach to retain flexibility for the provider to source expertise as appropriate for their service (for example, taking into account any existing expertise they have in-house), provided that the expert has experience in deterring CSEA offences.

## Benefits and effectiveness

### Benefits

9.101 General search services are one of the most common ways for perpetrators to locate and access CSAM online.[1458] Evidence from the NCA suggests that CSAM can be found within three clicks on mainstream search engines.[1459]

9.102 The chapter titled ‘Search services’ in the Register provides additional evidence demonstrating how search engines can act as a pathway to accessing CSAM. Two of our measures in other chapters aim to address this risk. In chapter 5 of this Volume: ‘Automated search moderation’ (‘ASM’), one automated search moderation measure aims to remove listed CSAM URLs from search results, while our search moderation measures in chapter 3: ‘Search moderation’ place expectations on search service providers to ensure they have effective systems and processes in place to take appropriate action against illegal search content, including CSAM. However, this will not eliminate the risk of search results returning CSAM entirely, nor remove the possibility of a user searching for this content.

9.103 This measure aims to address this risk by acting as a point of friction in the user journey towards encountering CSAM via general search services. In doing so, the measure:

- deters users from searching for CSAM; and
- provides supportive resources to potential perpetrators.

9.104 In terms of deterrence, we anticipate that informing users of the illegality of viewing CSAM via a warning message may deter them from engaging with search results that contain CSAM, or from attempting to conduct such searches in future, due to fear of the consequences. Where CSAM is available in more than one click, the warning message may also help to disrupt the user journey before a user reaches the point of viewing illegal content.

9.105 In terms of support, providing links to relevant resources may help potential perpetrators refrain from committing CSEA offences if they are seeking out CSAM.

9.106 The motivations and behaviour of CSEA perpetrators are not necessarily straightforward. We therefore expect that consulting with experts when developing the warning message will help service providers to tailor these messages more appropriately and signpost to appropriate resources, which would therefore improve the deterrence effect of the message.[1460]

[1458] Bailey, A., Allen, L., Stevens, E., Dervley, R., Findlater, D., and Wafers, S. Pathways and prevention for indecent images of children offending: A qualitative study, Sexual Offending: Theory, Research, and Prevention, 17. [accessed 18 November 2024].

[1459] See principle 6 (‘example’) in Home Office, 2020. Interim code of practice on online child sexual exploitation and abuse. [accessed 18 November 2024].

[1460] In the anecdotal example provided by the Lucy Faithfull Foundation discussed at paragraph 9.73, it noted that traffic to its support website returned at an increased rate after it collaborated with the search service provider to change its warning message. Source: Lucy Faithfull Foundation response to November 2023 Consultation, pp.6-7.

### Effectiveness

9.107 Current industry practice indicates that this measure is a technically feasible way for general search service providers to try to deter users from attempting to access CSAM via their services.[1461]

9.108 As BILETA indicated in its response, there is limited evidence for the effectiveness of content warnings in deterring users and directing them towards support.[1462] We recognise that this is a particularly difficult area to research, and that it can be challenging to evaluate the success of any particular messaging or intervention used to deter CSAM offending. That said, we have identified some supportive evidence from the available literature.

9.109 Evidence pointing to the benefits of search services signposting to support organisations in this way includes a 2014 study conducted on the effectiveness of the UK-based Stop It Now! helpline. This study found that 3.7% of callers discovered the helpline through search engines, and that those interviewed were supportive of the warning banners as a possible intervention to prevent their illegal behaviour.[1463] [1464] In a more recent evaluation of UK-based deterrence efforts, 2 of 55 callers to the Stop It Now! helpline during its 2023 deterrence campaign became aware of the service through 'splash pages' while 25 of 55 callers had become aware of the service through internet searches seeking support resources.[1465] We note that the actual number of perpetrators positively affected by the interventions in these studies is very small in comparison to the large number of people who use search services. Given the challenges of gathering evidence in this space, as outlined in paragraph 9.108, we think it is appropriate to view this as evidence of the effectiveness of this measure, not least as one perpetrator alone can cause severe harm.

9.110 There is also evidence relating to the Finland-based ReDirection programme website (an anonymous, online self-help resource), which had been visited 80,000 times within two years of its launch in September 2021. Most of these users accessed the programme's website after viewing intervention messages on dark web search engines.[1466] The report

[1461] Both Google Search and Microsoft Bing display content warnings in response to user searches for CSAM. Google Search's deterrence message includes information on how to report CSAM to the Internet Watch Foundation. It also provides a link to the Lucy Faithfull Foundation's 'Stop it Now!' campaign, which focuses on the prevention of child sexual abuse and offers a broad range of support (including for those that are worried about their own thoughts or behaviour). Source: Ofcom desk research, conducted 18 November 2024. Microsoft Bing's warning message also provides a link to the "Stop it Now!" campaign. The National Crime Agency (NCA) provided Microsoft with a list of keywords to trigger the message. Source: Microsoft UK Stories, 2013. Microsoft and Google stand united to combat online child sexual abuse content. [accessed 18 November 2024].

[1462] BILETA response to November 2023 Consultation, p.19.

[1463] Stop It Now! is a confidential helpline and online chat service run by the Lucy Faithfull Foundation for anyone with concerns about child sexual abuse, including those concerned by their own thoughts and behaviour. Source: Stop it Now. How we prevent child sexual abuse. [accessed 18 November 2024].

[1464] Brown, A., Jago, N., Kerr, J., McNaughton Nicolls, C., Paskell, C., and Webster, S., 2014. Call to keep children safe from sexual abuse: A study of the use and effects of the Stop it Now! UK and Ireland Helpline. [accessed 18 November 2024]

[1465] The Lucy Faithfull Foundation, 2023. Stop It Now! 2023 online CSA deterrence campaign (Phase 8) evaluation. A splash page is an introductory page or pop-up on a website. The 55 callers to the helpline were individuals who had not been arrested.

[1466] The report does not quantify how many visitors accessed the ReDIrection Programme website via the dark web. Source: Protect Children, 2023, 'Chat to a specialist': Evaluation of an anonymous chat function of the ReDirection program. [accessed 18 November 2024].

notes that while many visitors do not actively participate in the programme, those who open the first section of the resource tend to complete the programme.[1467] Of those users who went on to complete the ReDirection programme and provided feedback, 77% said that their use of CSAM had reduced or stopped completely.[1468]

9.111 Furthermore, we received an anecdotal example from the Lucy Faithfull Foundation demonstrating how displaying a warning message without signposting the user to relevant information can reduce referrals from search services to support organisations.[1469]

9.112 We note that some evidence suggests there are limitations to the effectiveness of CSAM warnings. A qualitative study of the role of pathways and prevention in reducing CSAM offending drew on data collected through interviews with 20 individuals arrested during 2015 and 2016 for CSAM offences.[1470] The majority of participants reported never encountering any online deterrence messages. The six respondents who reported encountering online deterrence messages said they did not find such messages to be effective and noted that they could be easily ignored. However, several participants cited other techniques they felt would have helped them avoid offending prior to arrest:

- Four participants felt that online and offline messages regarding the illegality of CSAM would have helped to prevent their use of it.
- Six participants felt that warnings about the consequences of viewing CSAM would have helped to prevent their use of it.
- Five participants felt that having knowledge of what help was available for those seeking out CSAM would have helped to prevent their use of it.[1471]

9.113 These responses suggest that warnings with the right components (for example, messages around the illegality of CSAM) may be effective at preventing and deterring CSAM offending.

9.114 Based on the dataset discussed in paragraph 9.112, the Lucy Faithfull Foundation's report on its experience of deterrence campaigning set out four components of an effective deterrence campaign message:

a) viewing CSAM is illegal;
b) this behaviour causes harm to children;
c) there are serious consequences to CSAM offending for the offender and their loved ones; and
d) there is confidential support available.[1472]

[1467] 1,422 visitors opened the first section of the programme between September 2021 and June 2022. Of those visitors, 97% continued onto the second section, and 73% continued to the third. Source: Protect Children, 2023, 'Chat to a specialist': Evaluation of an anonymous chat function of the ReDirection program. [accessed 18 November 2024].

[1468] Protect Children, 2023, 'Chat to a specialist': Evaluation of an anonymous chat function of the ReDirection program. [accessed 18 November 2024].

[1469] Lucy Faithfull Foundation response to November 2023 consultation, p.7.

[1470] Allen, L., Bailey, A., Dervley, R., Findlater, D., Stevens, E and Wefers, S., 2022. Pathways and Prevention for Indecent Images of Children Offending: A Qualitative Study, *Sexual Offending: Theory, Research, and Prevention*, 17. [accessed 18 November 2024].

[1471] The study uses the term 'Indecent Images of Children Offending' rather than CSAM.

[1472] Lucy Faithfull Foundation (Denis. D., Findlater. D., and Walsh. M.), 2023. Deterring online child sexual abuse and exploitation: lessons from seven years of campaigning. [accessed 18 November 2024].

9.115 Components (i) and (iv) are covered under the provisions of this measure. We have not made specific recommendations around components (ii) and (iii) as we consider that providing a warning on the illegality of CSAM sends a strong deterrence message to potential perpetrators, making it clear that their behaviour is harmful and will have consequences. Service providers may wish to make components (ii) and (iii) more explicit in their CSAM warnings if they, or the person with expertise consulted, consider it appropriate to do so (while ensuring that messaging remains clear and accessible). We will continue to analyse emerging evidence on deterrence messaging and may iterate on this measure in future based on this evidence, if appropriate.

9.116 Another anecdotal example from the Lucy Faithfull Foundation illustrated how placement of the message at the bottom of the search results page made it stand out less to users, suggesting this might limit its effectiveness. [1473] We consider that this example reinforces the expectation we included in our original proposal, that service providers should prominently display the CSAM warning in the search results. As noted at paragraph 9.73, since the November 2023 Consultation, the Lucy Faithfull Foundation has worked with this service to improve their warning, which now appears at the top of the search results page.[1474]

9.117 While we acknowledge that warning messages may not disrupt intentional searches in all cases, they may act as a deterrent to users purposefully seeking out CSAM by disrupting their user journey. By deterring potential perpetrators, and providing them with support to curb their behaviour, the warning message should reduce the number of perpetrators that view CSAM.

9.118 We know that CSEA victims and survivors can experience re-traumatisation and continued re-victimisation due to knowing about, or inadvertently seeing, images of their own abuse circulating online.[1475] By reducing searches for CSAM, the harm inflicted on victims and survivors by the subsequent re-viewing and re-sharing of this content should also reduce. This will help to reduce the re-traumatisation of victims and survivors of abuse.

9.119 If a user's encounter with a deterrence message disrupts their user journey to the extent that they stop viewing the search results and seek support (or even go on to reform their behaviour) this may disrupt their offender pathway from viewing CSAM to committing contact child sexual abuse, reducing the risk of this further harm to potential victims.[1476] This could lead to material benefits even if the number of potential perpetrators whose journeys are disrupted is small, due to the severity of the impact of CSAM.

9.120 In summary, through disrupting search journeys to deter potential perpetrators from accessing CSAM and providing support to help curb their behaviour, the measure will reduce the number of potential perpetrators that view CSAM. In turn, this may reduce the

[1473] Lucy Faithfull Foundation response to November 2023 Consultation, p.7.

[1474] Lucy Faithfull Foundation email dated 21 November 2024.

[1475] In a survey conducted by C3P, 69% of victims and survivors indicated that they worried constantly about being recognised, and almost a third (30%) had been identified online or in person by someone who had seen images of their abuse. Some victims and survivors reported being targeted and re-victimised by someone who had recognised them, including being propositioned or threatened. The sample consisted of 150 victims and survivors. Source: C3P, 2017. Survivors' Survey: executive summary 2017. [accessed 18 November 2024].

[1476] A Protect Children study found that of respondents who had viewed CSAM, 37% had previously sought direct contact with children after viewing CSAM. Source: Protect Children (Insoll, T., Ovaska, A., and Vaaranen-Valkonen, N.), 2021. CSAM users in the dark web. ReDirection Survey Report. [accessed 18 November 2024].

re-traumatisation of victims and survivors and reduce the risk that those potential perpetrators go on to commit contact child sexual abuse. These potentially significant benefits are in line with providers' duties to minimise the risk of users encountering illegal search content and to mitigate and manage the risk of harm to individuals under section 27 of the Act.

## Costs and risks

### Costs

9.121 We expect that the upfront costs to develop a warning system will include both the initial software development costs and developing a list of search terms related to CSAM (in response to which the warning message will be shown).

9.122 Implementing this measure may lead to initial costs for service providers that do not currently have a suitable warning system in place. They may also incur ongoing costs associated with maintaining and updating the system to ensure it functions correctly.

9.123 Software costs will depend on whether providers make use of regularly updated third-party lists or third-party solutions. If they use regularly updated third-party lists, they will need to adopt appropriate processes for copies of lists that they hold, including for integrating them into their systems and maintaining appropriate safety procedures. If providers opt for third-party solutions, they will need to ensure secure and efficient integration in terms of security and performance.

9.124 Service providers may incur costs from purchasing external lists, developing their own lists, or a combination of both. As outlined in paragraph 9.84, there are likely to be third-party organisations with existing lists that can offer these to providers. These will potentially be available under the same arrangements discussed in our measure on removing CSAM URLs from search results, as recommended in chapter 5 of this Volume: 'ASM'. This will likely reduce the implementation and running costs of this measure.

9.125 We expect that the software development needed to apply this measure will take between 170 to 310 days of software engineering time depending on the complexity and functioning of the system (along with an equal amount of time input from professional occupation staff). We estimate one-off implementation costs of around £80,000 to £310,000.[1477] The total implementation cost will depend on the complexity of the search system, how messages are displayed, the extent of identified search terms, and the labour costs assumed for software engineers and other professionals.[1478]

9.126 We expect annual maintenance costs to be equivalent to 25% (one-fourth) of the implementation cost and estimate this to be £21,000 to £80,000 per year.[1479] Ongoing

[1477] Based on our labour cost assumptions set out in Annex 5. We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS. We received some feedback on the general cost assumptions (such as salary assumptions) that are fed into these costs. We consider that feedback in Annex 5.

[1478] The professions we have determined to be most relevant for developing our proposed measures are described in Annex 5.

[1479] As described in Annex 5, we assume annual maintenance costs are 25% (one-fourth) of the initial costs where we have no more specific information. We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS. We received some feedback on the general cost assumptions (for example, salary assumptions) that are fed into these costs. We consider that feedback in Annex 5.

running costs will likely include regularly updating the terms list and other miscellaneous system maintenance costs.

9.127 Service providers will also need to identify reputable organisation(s) dedicated to tackling child sexual abuse that freely provide suitable resources. While this may require some research, we expect the costs of this to be minor because there are only a few relevant organisations accessible to UK users and the task of creating appropriate warnings is not overly complex. We also expect there will be small ongoing costs involved in ensuring the warnings remain updated and current.

9.128 Service providers will also likely incur some additional costs in engaging with experts in deterring CSEA offences (internal or external) while developing their CSAM warning messages. These will include the costs of finding and consulting with such experts and taking their suggestions on board. These costs may be in addition to those incurred when engaging with expert(s) regarding the terms commonly used by CSAM, since those experts may not necessarily have specific expertise in deterring CSEA offences. On balance, we consider that these costs are proportionate given the potential benefits outlined in paragraph 9.103.

9.129 Both Google Search and Microsoft Bing already display content warnings in response to user searches for CSAM. We therefore expect that the providers of these services will incur negligible or limited additional costs when implementing this measure unless they intend to remove this feature.

9.130 Furthermore, search providers operating in Australia that are subject to the eSafety Search Code will be subject to requirements that are similar to those of this measure, and will therefore need to take actions similar to those recommended by this measure in any case.[1480] This may also limit the costs that providers need to incur to apply this measure.

### Risks

9.131 With this measure being applied only to providers of large general search services, it is possible that perpetrators will move to smaller services to search for CSAM. However, we consider this risk to be low because this measure is less likely to be effective in deterring users who are determined to search for CSAM material. We consider that such users are more likely to ignore any warnings given and proceed with their search (rather than moving to a different search engine).

## Rights impact

### Freedom of expression

9.132 Paragraph 9.50 explains the right to freedom of expression in Article 10 of the ECHR. It is not clear that the presentation of a warning message when a user enters search terms or combinations of letters and symbols that obviously relate to CSAM is an interference with the user's right to free expression. It may discourage engagement with search content but

[1480] Specifically, relevant search providers must "(g) ensure that search results specifically seeking images of known CSAM are accompanied by deterrent messaging that outlines the potential risk and criminality of accessing images of CSAM; and (h) ensure that search results returned for end-user queries using terms that have known associations to child sexual exploitation material (CSEM) are accompanied by information or links to services that assist Australian end-users to report CSEM to law enforcement and/or seek support" Source: eSafety, 2023. Schedule 6 - Internet Search Engine Services Online Safety Code (Class 1A and Class 1B Material), paragraphs 7(2)(g) and 7(2)(h). [accessed 18 November 2024].

will only be displayed in circumstances where there is a high likelihood that a user would be committing a crime if they viewed the content they were searching for. It would not prevent them from going on to view the content.

9.133 CSAM is an extremely harmful kind of illegal content. As such, even if this measure was an interference with the right to free expression, restrictions such as warnings that deter engagement are clearly justified insofar as they contribute to the prevention of crime, the protection of morals, and the protection of the rights of others (in particular, those of the child victims and survivors concerned).

9.134 For the reasons outlined in paragraphs 9.132 and 9.133, we do not consider the presentation of a CSAM warning message in line with this measure to have any material impact on the freedom of expression rights of either the providers of search services or website operators.

### Privacy

9.135 Depending on how service providers decide to implement this measure, it may result in a greater or lesser impact on users' privacy rights under Article 8 of the ECHR and their rights under data protection law as set out in 'Introduction, our duties, and navigating the Statement', and in chapter 14 of this Volume: 'Statutory tests'. We recognise that the implementation of this measure requires providers to analyse search queries to determine whether it is a query that should trigger the presentation of a CSAM warning message. In circumstances where a user enters a search query with the expectation that the provider will understand the request and present relevant search results in response to it, users can reasonably expect that the query inputted to the search bar will be processed to return search results (including a CSAM warning where relevant) which may include analysis of the query itself and other information that will enable the provider to present search results relevant to the user. For the purposes of this measure, providers are likely to also need to ascertain whether the user is located in the UK, in order to determine whether it needs to return the CSAM warning message. We consider that any interference with users' rights to privacy, including through the processes outlined above, is necessary to ensure providers fulfil their safety duties under the Act and is proportionate to the benefits to users and safeguarding of child victims who may suffer continued harm by users accessing CSAM.

### Data protection

9.136 This measure does not specify that service providers should obtain or retain any specific types of personal data about individual users as part of their implementation of this measure. However, as explained in paragraph 9.135, we acknowledge that the analysis of search requests to determine whether a warning should be displayed will involve processing the personal data of the user conducting the search. In particular, we recognise that this may involve the generation of criminal offence data about the user concerned.[1481] Providers choosing to process additional personal data in their analysis of search requests (or in any other activity involved in implementing this measure) will need to comply with relevant data protection legislation. This includes applying appropriate safeguards to protect any criminal offence data, and the rights of children (who may require special

[1481] A user who intentionally carries out a search for CSAM knowing that it will generate CSAM images on their computer screen is likely to be "making" a CSAM image (if they find such images) or attempting to "make" a CSAM image (if they don't).

consideration) and adults who will be affected by this measure. Providers should refer to relevant guidance from the ICO.[1482]

9.137 We therefore consider that the impact of this measure is likely to be limited where providers comply with relevant laws, and that any interference is both necessary to ensure compliance with the Act and proportionate.

## Who this measure applies to

9.138 In our November 2023 Consultation, we proposed recommending this measure to providers of large general search services. We considered the measure to be proportionate for such services as the costs are likely to be relatively small compared to the measure's potential benefits in reducing the risk of harm.

9.139 As discussed in paragraph 9.77, we received stakeholder feedback that the measure should apply to all services as the harm can also arise on smaller services.[1483]

9.140 However, we remain of the view that this measure will not be proportionate for smaller general search services for two main reasons.

- The smaller reach of these services makes it likely that the measure would affect fewer potential perpetrators. Therefore, the benefits of the measure may be limited for smaller general search services. In addition, we expect that the costs could still be material for a smaller general search service in some scenarios. Considering the mixed evidence of the effectiveness of this measure (as outlined in paragraph 9.112), we do not consider that the benefits would necessarily outweigh the costs for smaller services.
- The risk of displacement of users to smaller services in direct response to this measure is likely to be small. We consider that users who do not respond positively to the warning by ceasing to search for CSAM are more likely to ignore future warnings than move to a smaller search service that does not display warning messages.

9.141 We also have some concerns that a measure of this type (with a relatively fixed cost of implementation) could make it materially difficult for new entrants looking to enter the search market. This has also influenced our decision not to apply this measure to smaller services as we do not want to discourage competition from new entrants to the market.

9.142 We received stakeholder feedback that the measure should extend to search functions on services beyond large general search services, which could include both smaller search services, and the search functions on U2U services.[1484] We have not received sufficient evidence that changes our assessment of the proportionality of recommending this measure for smaller services at this stage (as set out at paragraph 9.140).

9.143 We are aware that some U2U services already implement similar measures. For example, Aylo (in collaboration with the Internet Watch Foundation (IWF) and the Lucy Faithfull Foundation) recently trialled a chatbot and warning message to disrupt searches for CSAM

[1482] ICO, UK GDPR guidance and resources and Online safety and data protection. [accessed 18 November 2024].

[1483] C3P response to November 2023 Consultation, p.29; Lucy Faithfull Foundation response to November 2023 Consultation, p.8; WeProtect Global Alliance response to November 2023 Consultation, p.26.

[1484] C3P response to November 2023 Consultation, p.29; Lucy Faithfull Foundation response to November 2023 Consultation, p.8; WeProtect Global Alliance response to November 2023 Consultation, p.26.

on its Pornhub service in the UK.[1485] While we were not in a position to recommend this for U2U services at the time of consultation, we welcome innovation by U2U services in this space and will keep this area under review for future iterations of our Codes as the evidence base grows.

9.144 We therefore consider that this measure is appropriate for large general search services.

## Conclusion

9.145 We consider this measure would have potentially significant benefits. While it will impose some costs, these are not disproportionately large given the important benefits the measure will deliver and the fact we are targeting it at large general search services. It is less clear that it would be proportionate to extend the scope of the measure to include smaller search services. Therefore, we have decided to apply this measure to large general search services.

9.146 This measure is included in our Illegal Content Codes of Practice for search services on CSEA. It is referred to within this Code as 'ICS F2'.

# Measure on provision of suicide crisis prevention information

9.147 In our November 2023 Consultation, we proposed that providers of large general search services should detect search requests that contain general queries regarding suicide, and queries seeking specific, practical, or instructive information regarding suicide methods. Providers should provide crisis prevention information in response to such search requests.[1486]

9.148 The aim of the proposed measure was to reduce the risk of users encountering content encouraging or assisting suicide in the search results and mitigate the risk of very serious harm to users in that context.

## Summary of stakeholder feedback[1487]

9.149 Three civil society organisations noted their support for the provision of suicide crisis prevention information in principle, while also suggesting several areas where the measure could be strengthened or clarified.[1488] These included (but were not limited to):

[1485] Internet Watch Foundation, 2024. Pioneering chatbot reduces searches for illegal sexual images of children. [accessed 18 November 2024]. As noted in Scottish Government response to November 2023 Consultation, p. 12, and Lucy Faithfull Foundation response to November 2023 Consultation, p.9.

[1486] In our November 2023 Consultation, we referred to this measure as Measure 7C. We have now renumbered it ICS F3.

[1487] Note this list is not exhaustive, and further responses can be found in Annex 1.

[1488] NSPCC response to November 2023 Consultation, p.45; Samaritans response to November 2023 Illegal Harms Consultation, p.3. 5Rights Foundation also called for Ofcom to consider the NSPCC's response on this crisis prevention measure. Source: 5Rights Foundation response to November 2023 Consultation, p.33. We received support for the crisis prevention measure (Measure SD2) in the draft Children's Safety Code. See CELCIS response to May 2024 Consultation, p.18; Jamie Dean response to May 2024 Consultation, p.20; Nexus response to May 2024 Consultation on Protecting Children from Harms Online, p.23; NSPCC response to May 2024 Consultation, p.72; Scottish Government response to May 2024 Consultation, p.19.

- clarification of the search query categories triggering crisis prevention information;[1489]
- how providers should stay up to date with evolving language;[1490]
- the content of the crisis prevention information;[1491]
- the relationship between service providers and third-party support organisations;[1492]
- our assessment of the impact of the measure on privacy;[1493] and
- applying the measure to all search services.[1494]

9.150 We outline these stakeholder concerns in more detail in the following sections.

## Categories of search queries that trigger crisis prevention information

9.151 Samaritans requested further clarity on what searches the category of "general queries regarding suicide" would include, for example, searches for online challenges or high-profile deaths by suicide. It argued that these should be included in the categories of search requests that this measure would cover.[1495]

9.152 Publicly available guidance by Samaritans around media reporting on celebrity suicides describes how high-profile deaths by suicide contribute to the phenomenon of "suicide contagion". It cites a 2020 study indicating that celebrity suicides are associated with a 13% increase in suicides in the subsequent one to two months.[1496] It explained that some individuals might "over-identify with celebrities" due to media exposure of information about their private lives, which may lead to imitative suicidal behaviour.

9.153 In its consultation response, Samaritans highlighted that some items commonly used for suicide have everyday uses, and that there is a risk of inadvertently raising awareness of suicide methods if crisis prevention information is displayed when search requests about such items are made with no suicidal intention behind the search. It commented that efforts to avoid this risk are not happening consistently across all search services.[1497]

9.154 We address these concerns in paragraphs 9.174 to 9.179 in the section entitled 'How this measure works'.

## Staying up to date with evolving language

9.155 The Molly Rose Foundation discussed the challenge of online communities using deliberately obscure search terms. It argued that service providers should have "ongoing

[1489] Samaritans response to November 2023 Consultation, p.5.

[1490] Molly Rose Foundation response to November 2023 Illegal Harms Consultation, p.39.

[1491] NSPCC response to November 2023 Consultation, p.45; Samaritans response to November 2023 Consultation, p.5; Scottish Government response to May 2024 Consultation, p.19.

[1492] NSPCC response to November 2023 Consultation, p.45.

[1493] ICO response to November 2023 Consultation, p.21.

[1494] C3P response to May 2024 Consultation, p.32; East Riding Safeguarding Children Partnership (ERSCP) response to May 2024 Consultation, p.4; Samaritans response to November 2023 Consultation, pp.3-4.

[1495] Samaritans response to November 2023 Consultation, p.5.

[1496] Samaritans, Guidance for reporting on celebrity suicides and suicide attempts. [accessed 18 November 2024]. Arendt. F., Braun. M., Cheng. Q., Niederkrotenthaler. T., Pirkis. J., Scherr. S., Sinyor. M., Spittal., Stack, S., Till. B., M., Tran. U. S., Voracek. M., and Yip. P. S. F., 2020. Association between suicide reporting in the media and suicide: systematic review and meta-analysis, *British Medical Journal*, 368. [accessed 18 November 2024].

[1497] Samaritans response to November 2023 Consultation, p.3.

detection and monitoring processes to track emerging changes in user behaviour and search terms", enabling them to take action accordingly.[1498]

9.156 We address this concern in paragraph 9.184 to 9.185 in the section entitled 'How this measure works'.

### Content of the crisis prevention information

9.157 Samaritans argued that any support services signposted as part of this measure should be 24/7 services due to the constantly available nature of the internet. It highlighted research suggesting that users purposefully browsing for suicide content may be in distress, therefore requiring access to immediate support.[1499] We address this concern in paragraphs 9.188 to 9.189 in the section entitled 'How this measure works'.

9.158 The National Society for the Prevention of Cruelty to Children (NSPCC), supported by the 5Rights Foundation, argued that suicide crisis prevention information should be appropriate for both children and adults, and should include helplines or organisations that children and young people will recognise.[1500]

9.159 We address these concerns in paragraph 9.190 in the section entitled 'How this measure works'.

### Relationship with third-party support organisations

9.160 The NSPCC highlighted the risk that support services might become overwhelmed by increased demand if a service provider began signposting to a support service without seeking consent from (or collaborating with) that organisation.[1501] In its response to the crisis prevention measure in the May 2024 Consultation, the NSPCC made the same suggestion and argued that seeking consent would place a "very limited burden" on search providers and allow for alignment with measure PCU E3 in the draft Childrens' Safety Code for U2U services.[1502] Measure PCU E3 proposed that providers should signpost children to support at key points in the user journey, and if any third-party organisation signposted to is not a public body, the provider should obtain the consent of that organisation to do so.[1503]

9.161 We note that Mid Size Platform Group expressed concern that obtaining consent from third-party organisations under the U2U signposting measure might lead those organisations to expect benefits from service providers in return for their consent, which could disadvantage smaller services.[1504]

[1498] Molly Rose Foundation response to November 2023 Consultation, p.39.

[1499] Samaritans response to November 2023 Consultation, p.5.

[1500] 5Rights Foundation response to November 2023 Consultation, p.33, NSPCC response to November 2023 Consultation, p.45.

[1501] NSPCC response to November 2023 Consultation, p.45; NSPCC response dated 12 June 2024 to our follow-up email dated 12 June 2024. The NSPCC's comments also applied to the measure relating to the provision of CSAM warnings. However, as we are not recommending that search services provide a helpline under that measure, we do not consider the risk of overburdening support services to be as relevant given there is no expectation for providers to signpost users to one-to-one support. While service providers might choose to direct users to a helpline, this would not be a direct outcome of the measure.

[1502] NSPCC response to May 2024 Consultation, p.70.

[1503] See 'PCU E3' in Annex 7 in May 2024 Consultation. [accessed 18 November]

[1504] Mid Size Platform Group response to May 2024 Consultation on Protecting Children from Harms Online, pp.12-13.

9.162 We address these concerns in paragraphs 9.208 to 9.210 in the section entitled 'Costs and risks'.

### Feedback on our rights assessment

9.163 As with the measure relating to CSAM warnings, the ICO recommended that we review our assessment of the impact of this measure on the right to privacy to ensure that it fully covered relevant data protection considerations.[1505] It argued that the measure could lead to providers processing personal data to deliver warnings to individual identifiable users (depending on how providers implement crisis prevention information).[1506] It also noted that the provision of crisis prevention information could require providers to process special category data relating to the health of users. We address these concerns in paragraph 9.217 to 9.219 under the section entitled 'Rights impact'.

### Services in scope of the measure

9.164 We received feedback from Samaritans that this measure should apply to both large and small search services, as several of the latter already provide crisis prevention information. It argued that we should focus on not disincentivising smaller services from continuing with the measure (rather than on the measure being a potential barrier to entry for new services).[1507] We received similar feedback from two stakeholders in response to our proposed measure around predictive search in the May 2024 Consultation.[1508]

9.165 We address these concerns in paragraphs 9.221 to 9.225 in the section entitled 'Who this measure applies to'.

## Our decision

9.166 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation. We have made some minor changes in response to the stakeholder feedback that we have received:

- We have recommended that the crisis prevention information includes a helpline offering a 24/7 service available to all UK users that is suitable for all ages.
- We have recommended that the crisis prevention information displayed by service providers should be comprehensible and suitable in tone and content for as many users as possible, including children.

9.167 Our measure now says that:

a) Providers of large general search services should employ means to detect and provide crisis prevention information in response to search requests made by users that contain general queries regarding suicide; and queries seeking specific, practical or instructive information regarding suicide methods. The crisis prevention information should provide a helpline that is both associated with a reputable health or suicide prevention organisation and is available to all UK users, irrespective of age or geographical location

[1505] ICO response to November 2023 Consultation, p.21.

[1506] ICO response to November 2023 Consultation, p.21.

[1507] Samaritans response to November 2023 Consultation, pp.3-4.

[1508] C3P response to May 2024 Consultation, p.32; East Riding Safeguarding Children Partnership (ERSCP) response to May 2024 Consultation, p.4.

within the UK, for 24 hours per day. It should also provide link(s) to information and support.

9.168 The full draft of the measure can be found in our Illegal Content Codes of Practice for search services for other duties and is referred to as measure ICS F3.

# Our reasoning

## How the measure works

9.169 Under this measure, service providers should seek to detect and provide crisis prevention information in response to search requests that fall within the following categories which we consider present a risk of leading to illegal content that encourages or assists suicide:[1509]

- general queries regarding suicide;
- queries seeking specific, practical, or instructive information regarding suicide methods.

9.170 We do not expect providers to include search terms that may be common among users experiencing thoughts of suicide but which do not have any apparent direct connection to suicide in the search request, such as terms relating to mood and anxiety symptoms, trauma, or negative life events.[1510]

### General queries regarding suicide

9.171 The first category in paragraph 9.169 ("general queries regarding suicide") intends to cover search requests that have a direct connection to suicide but are not related to a specific suicide method. In the November 2023 Consultation, we explained that this would include common suicide search terms such as "suicide" and "kill yourself".[1511]

9.172 The broad nature of this category could encompass both searches seeking help for suicidal ideation and searches relating to pop culture references to suicide.[1512] This is because this category aims to capture search requests that vulnerable users are likely to make when in an earlier, more speculative browsing phase in their search journey, provided the link to suicide is clear. However, as explained, we would not expect providers to cover terms that indirectly relate to suicide such as more general mood, anxiety or trauma symptoms.

9.173 The search requests covered by this category might not expressly lead to illegal suicide content on their own, but instead may lead vulnerable users down a path of increasingly harmful content that ultimately leads to searches directing them to content encouraging or assisting suicide. We discuss the benefits of disrupting search journeys at this more speculative stage in paragraphs 9.196 to 9.197.

9.174 We have carefully reviewed suggestions and evidence from Samaritans regarding searches about high-profile deaths by suicide and online challenges. We considered how they relate

[1509] Schedule 7, section 1-2 of the Act.

[1510] Ali. A., Birnbaum, M. L., Kane, J. M., Kirschenbaum. M. A., Moon. K. C., and Van Meter. A. R, 2021, Internet Search Activity of Young People With Mood Disorders Who Are Hospitalized for Suicidal Thoughts and Behaviors: Qualitative Study of Google Search Activity. *JMIR Mental Health,* 8(10). [accessed 18 November 2024].

[1511] Borge, O., Cosgrove, V., Cryst, E., Grossman, S., Perkins, S., and Van Meter, A., 2021. How Search Engines Handle Suicide Queries, *Journal of Online Trust and Safety*, 1(1). [accessed 18 November 2024].

[1512] "Suicidal ideation means thinking about or planning suicide. Thoughts can range from a detailed plan to a fleeting consideration." Source: NHS. Supporting someone with suicidal thoughts. [accessed 18 November 2024].

to our policy intention and to what extent these kinds of search requests fall within the categories of content covered by this measure (as outlined in paragraph 9.169).[1513]

9.175 There is substantial evidence demonstrating that the risk of 'suicide contagion' can arise in the context of media reporting where reports feature the suicide of a celebrity or high-profile person.[1514] As discussed in the Register chapter titled 'Encouraging or assisting suicide (or attempted suicide),' the contagion effect could also apply to online contexts. We recognise that exposure to content about a high-profile death by suicide in the search results could have potentially serious consequences for vulnerable users if they were to overly identify with a high-profile victim of suicide when encountering this content. However, we have not seen evidence that searches for high profile deaths by suicide in particular present a risk of users encountering illegal content over and above other types of searches.

9.176 Depending on their precise wording and the context in which they were made, we consider that some searches for high-profile deaths by suicide would fall into the 'general queries regarding suicide' category or the 'suicide methods' category, as described at paragraph 9.169. However, beyond this we do not consider it necessary to prescribe that all searches related to a high-profile death by suicide should surface crisis prevention information. We are concerned that expanding the scope of the category in this way could increase the risk of desensitising users if providers display crisis prevention information too frequently where there is limited risk of harm to users. We discuss the risk of desensitising users in more detail in paragraph 9.211.

9.177 We have therefore decided not to amend the measure to explicitly include searches for high-profile deaths by suicide (as requested by Samaritans).[1515]

9.178 For similar reasons, we do not consider it appropriate for all search requests citing online challenges (such as requests including the name of a challenge that might result in accidental death rather than a specific suicide challenge) to be within the scope of this measure. However, a query which specifically mentions a suicide method involved in an online challenge may be covered by "queries seeking specific, practical, or instructive information regarding suicide methods." In this instance it would be the mention of the method, rather than the challenge itself, that would be the relevant factor.

9.179 Beyond the circumstances outlined in paragraph 9.178, we are not recommending that service providers display crisis prevention information in response to queries around dangerous challenges that might result in accidental death. We consider it unlikely that these search requests will direct users to suicide content within the scope of the illegal content safety duties. Such challenges are more directly captured by the child safety duties that apply to search services under section 29 of the Act, as "content which encourages, promotes, or provides instructions for a challenge or stunt highly likely to result in serious injury to the person who does it or to someone else" is a category of priority content

[1513] Samaritans response to November 2023 Consultation, p.5.

[1514] Samaritans, Guidance for reporting on celebrity suicides and suicide attempts. [accessed 18 November 2024]. Arendt. F., Braun. M., Cheng. Q., Niederkrotenthaler. T., Pirkis. J., Scherr. S., Sinyor. M., Spittal., Stack, S., Till. B., M., Tran. U. S., Voracek. M., and Yip. P. S. F., 2020. Association between suicide reporting in the media and suicide: systematic review and meta-analysis, *British Medical Journal*, 368. [accessed 18 November 2024].

[1515] Samaritans response to November 2023 consultation, p.5.

harmful to children.[1516] The search moderation measures and equivalent predictive search reporting measure in the draft Children's Safety Codes of Practice for search services will contribute to minimising the risk of children encountering this content and experiencing the harms associated with it.[1517]

### Queries seeking specific, practical or instructive information regarding suicide methods

9.180 The second category outlined in paragraph 9.169 ("queries seeking specific, practical, or instructive information regarding suicide methods") may encompass some searches for instructions or resources about the experience of using one of those methods.

9.181 When designing this measure, we have been aware of the risk of unintended consequences, including leading vulnerable people to become aware of suicide methods (as noted by Samaritans in paragraph 9.153).[1518] We note evidence of this happening in other contexts, leading to the increased use of those methods. Outside the UK, for example, there have been instances where the use of a novel suicide method has increased after the method was reported by traditional media.[1519]

9.182 We consider that our emphasis on search requests that seek "specific, practical, or instructive information" will help drive consistent awareness of this risk with general search service providers. Under this category, we would not expect a clearly neutral query about a suicide method with an everyday use to trigger crisis prevention information. Instead, we expect providers to display crisis prevention information in response to queries relating to suicide methods that they consider are likely to direct users to content assisting or encouraging suicide (such as where it is clear that the user is seeking practical instructions on how to use the method in question to take their life).

9.183 We maintain that providers are best placed to decide how to identify related search requests and to determine which search terms or requests should trigger the display of crisis prevention information within the categories outlined in paragraph 9.169.

9.184 We recognise that search service providers who fail to stay up to date with evolving language in this space may be unable to provide crisis prevention information in response to high-risk searches using newer terminology. As a result, a vulnerable user searching for suicide-related content could be more likely to encounter illegal content encouraging or assisting suicide and less likely to receive timely assistance to mitigate the risk of harm. As highlighted by the Molly Rose Foundation, our 2024 research on the Prevalence of Non-Suicidal Self-Injury, Suicide, and Eating Disorder Content Accessible by Search Engines demonstrates that the language used by vulnerable communities can evolve as users seek to avoid detection.[1520] It is reasonable, then, to assume that search queries may evolve as novel methods of suicide emerge.[1521]

[1516] Section 62(8) of the Act. We provide more information on this harm in Section 8.10 of Ofcom's draft Guidance on Content Harmful to Children in the May 2024 consultation. [accessed 18 November 2024].

[1517] See Annex 8 in the May 2024 consultation. [accessed 18 November 2024].

[1518] Samaritans response to November 2023 consultation, p.3.

[1519] Chen. F., Cheng. Q., and Yip. P. S. F., 2017. Media effects of suicide methods: A case study on Hong Kong 1998-2005, *PLoS One*, 12(4), [accessed 18 November 2024].

[1520] Molly Rose Foundation response to November 2023 Consultation, p.39.

[1521] Network Contagion Research Institute, 2024. One Click Away: A Study on the Prevalence of Non-Suicidal Self Injury, Suicide, and Eating Disorder Content Accessible by Search Engines. [accessed 18 November 2024].

9.185 We would therefore expect service providers to keep the search terms triggering the display of crisis prevention information under review to ensure that they effectively cover the categories outlined in paragraph 9.169. This should be the case whether a service provider is using a keyword matching approach, or a more sophisticated method using machine learning, for example. This will help to ensure that the crisis prevention information reaches as many users as possible who are searching for illegal content encouraging or assisting suicide, whether they are using established or new terms relating to suicide and suicide methods.

### Content of crisis prevention information

9.186 Under this measure, service providers detecting applicable search queries as described in paragraph 9.169 should display links to freely available information provided by reputable mental health organisations and a helpline associated with such an organisation.

9.187 Crisis prevention information can be displayed in several ways, including by prioritising crisis prevention services in the search results or by providing crisis prevention information in a banner. Providers may choose to offer this information in any format they consider appropriate if it is prominently displayed to users.

9.188 Support should be available to users in distress in a timely manner that disrupts their search journey, to both reduce their risk of encountering content encouraging or assisting suicide and increase the likelihood that they access the support available. In a September 2021 report, Mental Health Innovations found that, of the 27,600 students who had sought support for their mental health from the Shout texting service so far that year, 75% had contacted the service outside the hours of 9am to 5pm. This highlights the importance of ensuring that support is continually available.[1522] While two large general search service providers (Google and Microsoft) already signpost users to 24/7 helplines in their suicide crisis prevention information, we are concerned users could be negatively affected if a service provider decided to stop this practice. We have therefore decided to amend this measure to recommend that the crisis prevention information includes a helpline offering a 24/7 service available to all UK users that is suitable for all ages.[1523] Service providers will still have the flexibility to signpost users to additional non-24/7 helplines if they wish to do so.

9.189 In making the recommendation given in paragraph 9.188, we acknowledge feedback we received around the risk of overwhelming support organisation helplines (see paragraph 9.160). We recognise that recommending that service providers list a 24/7 helpline may create a narrow pool of appropriate resources to which they can direct users. However, as the existing practice of both Google Search and Microsoft Bing is to signpost users to 24/7 services, we do not expect this new expectation to create any additional pressure for the support organisations providing those helplines.

9.190 We acknowledge that crisis prevention information should be appropriate for both children and adults to ensure it is effective for as many users as possible. We have therefore decided

[1522] Twenty percent texted Shout between 10pm-12am, 15% between 8pm-10pm and 13% between 12am-2am. Source: Mental Health Innovations, 2021. Supporting student mental health. Insight into students seeking support. [accessed 18 November 2024].

[1523] On conducting desk research into the availability of 24/7 helplines, we found that several of these support services have upper age limits. We therefore considered it important to specify that the helpline should provide a service available to UK users that is suitable to all ages to ensure that support is available to as many users as possible. 24/7 services operating in the UK include Samaritans, Shout, and NHS 111.

to amend this measure to recommend that the crisis prevention information displayed by service providers should be comprehensible and suitable in tone and content for as many users as possible, including children. This will also provide regulatory alignment with the crisis prevention measure proposed in the May 2024 Consultation.[1524]

9.191 We are not prescriptive about how the crisis prevention information should be provided, and therefore, at this stage we do not think it is appropriate to specify particular ways in which it should be accessible to disabled people. However, providers should consider their obligations under other relevant legislation (for example, the Equality Act 2010) and, where relevant appropriate guidance.[1525]

## Benefits and effectiveness

### Benefits

9.192 Search services are a gateway to information about suicide that exists online. Where suicide content intentionally encourages or assists a person to end their life, this may amount to a priority offence of encouraging or assisting suicide.

9.193 Research indicates content that celebrates, glorifies, or instructs users on self-injurious behaviour (including suicide) can be accessed within a single click from the main search results page.[1526] While not all search results of this nature will amount to the priority offence of encouraging or assisting suicide, the line between legal and illegal in this context is very difficult to draw and we therefore consider the findings of this research to be evidence of a clear risk.

9.194 The chapter titled 'Search services' in the Register sets out further evidence of the role of search services – and specific functionalities they provide to users – in making content that could amount to the priority offence of encouraging or assisting suicide available to users who are actively, speculatively, or even unintentionally searching for it.[1527]

9.195 There is also some evidence to suggest that behaviour on search services moves from periods of speculative browsing to specific and purposeful searches on methods of harm as suicidal intent increases.[1528]

9.196 If suicide crisis prevention information is the first information that a user encounters in response to a search request relating to suicide within the categories outlined in paragraph 9.169, this may effectively disrupt a search journey that might otherwise have led them to encounter illegal content amounting to the offence of encouraging or assisting suicide. While this measure may not remove the risk of encountering illegal content entirely (as a

[1524] PCS E2 in Annex 8 in the May 2024 Consultation. [accessed 18 November 2024].

[1525] For example, see the Web Accessibility Initiative. WCAG 2 Overview. [accessed 18 November 2024].

[1526] The research by the Network Contagion Research Institute (NCRI) found that 22% of the 37,647 individual search results links they assessed across five search engine services contained content that celebrates, glorifies, or instructs self-injurious behaviour within a single click from the main search results page. The report defined self-injurious behaviour as non-suicidal self-injury, suicide, and eating disorders. The research found that 1,580 links were likely to be in scope (promoting self-injury) of extreme (encouraging others to engage in self-injurious behaviour). Source: Network Contagion Research Institute, 2024. One Click Away: A Study on the Prevalence of Non-Suicidal Self Injury, Suicide, and Eating Disorder Content Accessible by Search Engines. [accessed 18 November 2024].

[1527] Most research on the topic of suicide is not specifically directed at "illegal content" as defined in the Act but at the harm itself, so may include both legal and illegal content.

[1528] Borge, O., Cosgrove, V., Cryst, E., Grossman, S., Perkins, S., and Van Meter, A., 2021. How Search Engines Handle Suicide Queries, Journal of Online Trust and Safety, 1(1). [accessed 18 November 2024].

user can ignore the warning and such content may remain accessible in the search results despite moderation), it effectively supports compliance with the duty to minimise the risk of individuals encountering priority illegal search content under section 27(3) of the Act.

9.197 Providing supportive information and signposting to a 24/7 helpline that is available to UK users and is suitable for all ages will help users in distress or crisis to seek timely assistance. This combination of information supports providers to comply with their duty to effectively mitigate and manage the risks of harm to individuals under section 27(2) of the Act. This benefit might occur immediately (for example, if the user in question accesses crisis prevention information resources when their search journey is disrupted) or it may minimise the risk of harm through the user remembering the content of the message or returning to it later.

9.198 We consider this measure to have significant potential benefits for users given the potential severity of the risk of harm to individuals in distress who encounter content encouraging or assisting suicide. We have therefore sought to ensure that this measure covers both those users who may be searching for content at an earlier or speculative phase of browsing (through the "general queries regarding suicide" category) and those who may be browsing purposefully for information about suicide methods while in an extremely vulnerable state (through the "queries seeking specific, practical, or instructive information regarding suicide methods" category).[1529]

### Effectiveness

9.199 Current practice indicates that providing suicide crisis prevention information in response to search requests related to suicide is a technically feasible way for providers to minimise the risk of users encountering illegal suicide content on their services and mitigate the risk of harm to individuals by providing timely assistance.[1530] [1531] [1532]

[1529] Research considering the search history of a sample of individuals hospitalised for suicidal thoughts and behaviour found that in 21% of cases, participants had searched for information that matched their chosen method of suicide. Source: Ali. A., Birnbaum, M. L., Kane, J. M., Kirschenbaum. M. A., Moon. K. C., and Van Meter. A. R, 2021, Internet Search Activity of Young People With Mood Disorders Who Are Hospitalized for Suicidal Thoughts and Behaviors: Qualitative Study of Google Search Activity. *JMIR Mental Health,* 8(10). [accessed 18 November 2024]. This indicates the benefits of providing crisis prevention information at the point of conducting this more specific category of search request to prevent users from encountering search content that encourages or assists suicide when in an extremely vulnerable state.

[1530] Google aims to provide suicide crisis prevention information where users in the UK express "urgent intent" around suicide. Source: Google response to the 2023 Ofcom Call for Evidence: Second phase of online safety regulation, p.49. It partners with crisis support services to display a feature at the top of the search results which includes Samaritans' helpline number and a link to its official website along with the facility to make a phone call via the mobile browser. Source: Samaritans, 2010. Press release: Google and Samaritans: new search feature to help people looking online for information about suicide. [accessed 18 November 2024], and Ofcom desk research, conducted 5 August 2024. Alongside this, it provides the number for Shout's text service and a link to its official website.

[1531] Microsoft aims to provide links to suicide prevention information resources where queries express "a possible suicide intent." It provides Samaritans' helpline number and a link to its website, alongside the number for emergency services, a link to the Campaign Against Living Miserably's support page, and a depression screening test from Mental Health America. Source: Microsoft Support, How Bing delivers search results. [accessed 18 November 2024].

[1532] Some smaller general search services (including DuckDuckGo, Ecosia, AOL and Yahoo) display crisis prevention information in response to search requests that include terms relating to suicide. Source: Ofcom desk research, conducted 18 November 2024.

9.200 Several charities, including those operating in the mental health and suicide prevention space, have endorsed the use of suicide crisis prevention information to reduce the risks of harm to vulnerable individuals. In response to Google's launch of its crisis prevention efforts on Google Search, Samaritans highlighted the importance of ensuring that "vulnerable and distressed people are steered towards safe spaces."[1533] Mental Health Innovations indicated that 2% of its daily conversations on the Shout support service were referred via signposts on Google Search and suggested that this demonstrates that "interventions such as this work to divert internet users" from potentially harmful searches.[1534]

9.201 Evidence indicates that prominently displaying crisis prevention information increases its effectiveness. A 2019 National Library of Medicine report found that higher ranked search results for suicide-related search requests (which were neutral and interspersed with anti-suicide pages) were more likely to be clicked on by users. The report concluded that efforts should be made to improve the visibility and ranking of suicide prevention webpages.[1535] This research suggests that users may be more likely to access crisis prevention information if it is highly ranked and that prominently displayed support information and helplines can help to minimise the risk of harm to users.

9.202 In summary, while providing crisis prevention information under this measure will not completely remove the risk of users encountering content that encourages or assists suicide, we maintain that it is likely to assist in safeguarding users through:

- effectively disrupting user search journeys to minimise the risk of individuals encountering illegal suicide content, in line with the duty under section 27(3) of the Act;
- providing timely assistance to individuals in distress where the risk of harm might otherwise be severe, in line with the duty to manage the risk of harm to individuals under section 27(2) of the Act.

## Costs and risks

### Costs

9.203 There will be an initial cost for providers that do not currently have this measure in place for their search services, along with ongoing costs to maintain and update the system to ensure it operates correctly. We expect the implementation costs will be moderate for service providers who already have a system in place that can provide information in response to specific search terms, as this will require a modification of the existing system to ensure it covers terms related to suicide.

9.204 We expect that implementing a new functionality and capability of this nature will require approximately 150 to 310 days of software engineering time (along with an equal amount of time input from professional occupation staff). We estimate this to be approximately £74,000 to £308,000 in one-off implementation costs (including the cost to develop an

[1533] Samaritans, 2010. Press release: Google and Samaritans: new search feature to help people looking online for information about suicide. [accessed 18 November 2024].

[1534] Mental Health Innovations UK response to 2023 Ofcom Call for Evidence: Second phase of online safety regulation, p.9.

[1535] Cheng. Q. and Yom-Tov. E., 2019. Do Search Engine Helpline Notices Aid in Preventing Suicide? Analysis of Archival Data, *Journal of Medical Internet Research*, 21(3). [accessed 18 November 2024].

interstitial displaying crisis prevention information).[1536] The total implementation cost will depend on the complexity of the search system, how messages are displayed, the extent of identified search terms, and the labour costs assumed for software engineers and other professionals.

9.205 We expect annual maintenance costs to be 25% (one-fourth) of the implementation cost, which is equivalent to £19,000 to £77,000 per annum.[1537] These running costs are likely to cover system maintenance and the costs of updating the system to ensure it properly identifies search requests related to suicide.

9.206 To implement this measure, service providers will need to identify reputable charities with 24/7 helplines that are available to UK users of all ages and provide access to resources related to suicide. While this may entail some research, we expect costs to be minimal as there are only a few relevant organisations accessible to UK users of all ages. We also expect that there will be small ongoing costs to ensure helplines and resource information remains updated and current.

9.207 Both Google Search and Microsoft Bing already provide suicide crisis prevention information, along with several smaller general search service providers. We expect these providers to incur negligible or limited additional costs when implementing this measure unless they intend to remove this feature. Also, the fact that both large and several smaller general search service providers already provide suicide crisis prevention information suggests that in practice the costs for this measure will not be excessive, at least for large general search service providers.

### Risks

9.208 We recognise that services within the scope of this measure have a large user base and have considered that this could result in an unmanageable number of users being directed to crisis prevention resources and services (in particular, 24/7 helplines). There is evidence that some viral user-generated content shared on a U2U service resulted in a spike of demand for Mental Health Innovations' Shout helpline. However, we are not aware of further evidence of support services becoming overwhelmed due to additional demand generated by crisis prevention efforts by search service providers.[1538] While the NSPCC have expressed concerns regarding this possibility, this measure would not contribute to a significant increase in traffic in practice.[1539] This is because two of the existing large general search services to whom the measure applies already signpost users to helplines provided by reputable mental health organisations (including 24/7 helplines) in their suicide crisis prevention information.

[1536] Based on our labour cost assumptions set out in Annex 5. We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS. We received some feedback on the general cost assumptions (for example, salary assumptions) that are fed into these costs. We consider that feedback in Annex 5.

[1537] As described in Annex 5, we assume annual maintenance costs are 25% (one-fourth) of the initial costs where we have no more specific information. We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS. We received some feedback on the general cost assumptions (for example, salary assumptions) that are fed into these costs. We consider that feedback in Annex 5.

[1538] Mental Health Innovation UK response to 2023 Ofcom Call for Evidence: Second phase of online safety regulation, pp.4-5.

[1539] NSPCC response to November 2023 consultation, p.45.

9.209 It is also unclear what costs and implications might be involved to obtain organisations' consent to list them in crisis prevention information (for example, whether there would be an expectation on the service provider to offer payment to the third party, as noted by Mid Size Platform Group at paragraph 9.161).[1540] It is not clear that the benefits of recommending that service providers seek the consent of the support organisations they signpost to would outweigh the potential costs of doing so.

9.210 We recognise that the benefits of seeking the consent of support organisations when signposting to them may be clearer where a service provider is seeking to make a change which is likely to increase traffic to a particular helpline (such as by removing or adding a helpline number to its crisis prevention information). While this would not be a direct impact of our measure, we encourage service providers to consider informing the relevant organisation where such a scenario may occur, allowing the organisation to plan for increased demand or flag if it is at capacity.

### Risks

9.211 There is a risk that displaying crisis prevention information too frequently might lead users to become desensitised to it (and therefore more likely to ignore it). Our research into user attitudes towards different types of interventions on services showed that participants became increasingly apathetic towards interventions as exposure increased over time.[1541] We expect this risk to be low as we have only recommended that service providers display crisis prevention information in response to the two categories outlined in paragraph 9.169.

## Rights impact

### Freedom of expression and freedom of association

9.212 Paragraph 9.50 explains the right to freedom of expression in Article 10 of the ECHR.

9.213 By providing supportive information at a critical point in the user journey, this measure disrupts pathways to illegal content that may encourage or assist suicide. It seeks to address the potentially very severe harm that users might experience if they actively search for this content or encounter it inadvertently by means of a search service. This is in line with the legitimate aims of the Act and the safety duties it imposes on search services.

9.214 This measure will not affect the search results displayed to users after crisis prevention information is presented to them. While it may introduce some friction into user journeys, the measure will not prevent users scrolling beyond the crisis prevention information and engaging with search results should they wish to do so. Therefore, we consider this measure to be one of the least restrictive ways to secure the objectives of the Act. We consider if there is any impact on the rights to freedom of expression of users, website operators, or search service providers, it is proportionate and justified in achieving the legitimate objectives of the Act.

9.215 There may also be a potential impact on freedom of association, as the presentation of crisis prevention information may deter users from encountering search results that would enable them to connect with other individuals who might be seeking support in connection with suicide. However, for the reasons outlined in paragraph 9.214, we consider that any

[1540] Mid Size Platform Group response to May 2024 Consultation, pp.12-13.

[1541] YouGov, 2023. User Attitudes towards On-Platform Interventions. [accessed 18 November 2024].

impact on the right to freedom of association of users is proportionate and justified in achieving the objectives of the Act.

### Privacy

9.216 Depending on how service providers decide to implement the measure, it may result in a greater or lesser impact on users' privacy rights under Article 8 of the ECHR as set out in 'Introduction, our duties, and navigating the Statement', and in chapter 14 of this Volume: 'Statutory tests'. We recognise that the implementation of this measure requires providers to analyse search queries to determine whether it is a query that should trigger the presentation of crisis prevention information. In circumstances where a user enters a search query with the expectation that the provider will understand the request and present relevant search results in response to it, users can reasonably expect that the data input to the search bar will be processed to return results (including crisis prevention information where relevant). This may also involve analysis of other information to enable the provider to present results relevant to the user. For the purposes of this measure, providers may also need to ascertain whether the user is located in the UK, in order to determine whether it needs to return the crisis prevention information and refer to UK-based resources and a UK-based helpline. We consider that any interference with users' rights to privacy, including through the processes outlined above, is necessary to ensure providers fulfil their safety duties under the Act and is proportionate to the benefits outlined in this chapter.

### Data protection

9.217 The measure does not specify that providers should obtain or retain any specific types of personal data about individual users as part of their implementation of this measure. However, we recognise that the analysis of search requests will involve processing personal data of the user conducting the search, some of which may be special category data in relation to a person's mental health (although this may be no more than would normally be processed in delivering search results).

9.218 Providers choosing to process additional personal data in their analysis of search requests (or in any other activity involved in implementing this measure) will need to comply with relevant data protection legislation. This would include applying appropriate safeguards to protect special category data, and the rights of children (who may require special consideration) and adults who will be affected by this measure. Providers should refer to relevant guidance from the ICO.[1542]

9.219 We therefore consider that the impact of this measure is likely to be limited where providers comply with relevant laws, and that any interference is both necessary to ensure compliance with the Act and proportionate.

## Who this measure applies to

9.220 In our November 2023 Consultation, we proposed to recommend this measure for providers of large general search services. We considered that this measure is likely to be proportionate for large services, given its likely benefits in reducing the risk of harm related to illegal suicide content and considering the costs outlined in paragraphs 9.203 to 9.207.

[1542] ICO, UK GDPR guidance and resources and Online safety and data protection. [accessed 18 November 2024].

We also acknowledged that Google Search and Microsoft Bing are already providing information of this type in response to search requests related to suicide.

9.221 As discussed in paragraph 9.164, we received stakeholder feedback that the measure should apply to both large and small service providers as several of the latter already provide crisis prevention information.[1543]

9.222 However, in our view, stakeholder feedback on who the measure applies to did not offer sufficient evidence to support recommending the measure for smaller services at this stage and we remain of the view that this measure will not be proportionate for smaller general search services for the following reasons:[1544]

- The benefits of applying this measure to a service with limited reach are likely to be relatively small, as the lower reach of smaller services means it is likely to disrupt fewer user journeys.
- Our analysis suggests that the implementation of this measure for smaller services could result in material costs in some scenarios that would not be proportionate.
- We also have some concerns that a measure of this type (with a relatively fixed cost of implementation) could make it materially difficult for new entrants looking to enter the search market. This has also influenced our decision not to recommend this measure for smaller services as we do not want to discourage competition from new entrants to the market.

9.223 We consider it unlikely that not including all services within the scope of this measure will lead smaller services who already provide crisis prevention information in response to applicable search requests to stop doing so. There will be existing reasons (such as user expectations) for these services to provide crisis information and we do not expect these reasons to change.

9.224 We note that providers of several smaller general search service services (such as DuckDuckGo, Ecosia, AOL, and Yahoo) already voluntarily provide crisis information of this type. We encourage them to continue to do so, notwithstanding the fact that we are not including them in the scope of the measure at this time.

9.225 We therefore continue to recommend this measure for large general search services.

## Conclusion

9.226 In light of this analysis, we conclude that this measure will deliver potentially significant benefits to UK internet users. While it will impose some costs on services, these are not disproportionately large given the scale of the benefits and the fact that we are targeting the measure at large general search services only. At the same time, our analysis shows that the impact the measure will have on privacy and freedom of expression rights is proportionate and justified. We have therefore decided to maintain our recommendation that the measure should apply to large general search services.

[1543] C3P response to May 2024 Consultation, p.32; East Riding Safeguarding Children Partnership (ERSCP) response to May 2024 Consultation on Protecting Children from Harms Online, p.4; Samaritans response to November 2023 Consultation, pp.3-4.

[1544] C3P response to May 2024 Consultation, p.32; ERSCP response to May 2024 Consultation, p.4; Samaritans response to November 2023 Consultation, pp.3-4.

9.227 This measure is included in our Illegal Content Codes of Practice for search services for other duties. It is referred to within this Code as 'ICS F3'.

# 10. Terms of service and publicly available statements

## What is this chapter about?

Terms of service ('terms') and publicly available statements ('statements') typically lay out the rights and responsibilities that a service provider and the users of their service have towards one another. Terms and statements are important to ensure transparency around the steps service providers are taking to protect users from illegal content. They are a tool for users to better understand the risk of using a service. This chapter sets out the measures relating to terms of service and publicly available statements we are recommending, why we are recommending them, and to which user-to-user (U2U) and search services they should apply.

## What decisions have we made?

We are recommending the following measures:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| ICU G1/ ICS G1 | Providers should include provisions in their terms and statements regarding **the protection of individuals** from illegal content, any **proactive technology used**, and information on **how complaints are handled and resolved**. | Providers of all services. |
| ICU G2/ ICS G2 | Providers should **summarise the findings** of their most recent **illegal content risk assessment** in their terms and statements. | Providers of Category 1 and Category 2A services. |
| ICU G3/ ICS G3 | Providers should ensure that provisions included in **terms and statements** regarding the protection of individuals from illegal content **are clear and accessible**. | Providers of all services. |

## Why are we making these decisions?

These decisions are intended to ensure that users understand the risks they face on relevant services and the measures service providers are taking to protect them from these risks. This will enable them to make more informed choices about what services to use. Not only will this allow users to better protect themselves from harm, but it may also generate reputational incentives for service providers to improve their safety measures.

## Introduction

10.1 As required by the Online Safety Act 2023 ('the Act'), service providers should design terms of service ('terms') and publicly available statements ('statements') for their regulated services which are both easy to access and understand. They must include provisions specifying how individuals are to be protected from illegal content, any proactive technology used, and provide users with information on how complaints are handled. These provisions must be consistently applied so that all users understand how they will be protected from illegal content.[1545]

10.2 In our November 2023 Illegal Harms Consultation ('November 2023 Consultation'), we proposed the following two measures as to how services can achieve their duties around terms and statements in accordance with the Act:

- The first proposed measure addressed how service providers should include in their terms and statements how they are protecting individuals from illegal content, any proactive technology used, and provide information on how complaints are handled and resolved. [1546]
- The second proposed measure recommended that service providers consider four factors when drafting provisions of their terms and statements: findability, layout and formatting, language, and usability.

10.3 Further, in our May 2024 Consultation on Protecting Children from Harms Online ('May 2024 Consultation'), we proposed an additional measure that providers of all Category 1 and Category 2A services develop or revise their terms and statements, ensuring they summarise the findings of their most recent illegal content assessment.[1547]

10.4 We received feedback from civil society organisations, service providers, and other stakeholders in response to our Consultations. We address the feedback directly relevant to each measure in the following sections. Other responses that we received are considered in Annex 1, including those submitted in relation to our use of prompts.

## Measure on substance of terms and statements

10.5 In our November 2023 Consultation, we proposed that all terms and statements include provisions in line with the duties in the Act: stating how the provider will protect individuals from illegal content, any proactive technology used, and information on how complaints are handled and resolved. We proposed this measure be applied to all user-to-user ('U2U') and search services.

[1545] Sections 10 (5), 10 (6), 10 (7), 10 (8) and 21 (3) for U2U services, and 27(5), 27 (6), 27(7), 27 (8) and 32(3) for search services, of the Online Safety Act 2023.

[1546] For this purpose, 'proactive technology' means (a) content identification technology, (b) user profiling technology, or (c) behaviour identification technology, as defined in Section 231 of the Online Safety Act 2023.

[1547] Some services will be categorised as Category 1, 2A or 2B if they meet certain thresholds set out in secondary legislation by Government. These categorised services will be required to comply with additional requirements. Category 1 services are regulated U2U services that meet the Category 1 threshold conditions. Category 2A services are search services that meet the Category 2 threshold conditions. Category 2B services are regulated U2U services that meet the Category 2B threshold conditions. For further detail see Ofcom, 2024. Categorisation: Advice Submitted to the Secretary of State [accessed 29 November 2024].

## Summary of stakeholder feedback[1548]

10.6 Most respondents were supportive of the measure.[1549]

10.7 A number of civil society organisations suggested that we should be more prescriptive about what service providers should include in their terms and statements, and offered suggestions of the types of additional information they could contain.[1550] These suggestions included:

- requiring service providers to supply clearer definitions and explanations of the various types of violating content;[1551]
- information about the types of enforcement action providers will take and in what circumstances;[1552] and
- the potential risk of harm to users from priority illegal content.[1553]

10.8 We address these suggestions in the 'Benefits and effectiveness' section.

[1548] Note this list is not exhaustive – further responses can be found in Annex 1.

[1549] Are, C. response to November 2023 Illegal Harms Consultation, p.13; Betting and Gaming Council response to November 2023 Illegal Harms Consultation, p.10; British and Irish Law Education and Technology Association response to November 2023 Illegal Harms Consultation, p.13; Canadian Centre for Child Protection (C3P) response to November 2023 Illegal Harms Consultation, p.23; Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) response to November 2023 Illegal Harms Consultation, p.11; Name withheld 5 response to November 2023 Illegal Harms Consultation, p.13; Dean, J. response to May 2024 Consultation on Protecting Children from Harms Online, p.18; Dwyer, D. response to November 2023 Illegal Harms Consultation, p.7; Evri response to November 2023 Illegal Harms Consultation, p.7; Federation of Small Businesses response to November 2023 Illegal Harms Consultation, p.4. We note that Federation of Small Businesses made a similar point in May 2024 Consultation on Protecting Children from Harms Online, p.7; 5Rights Foundation response to November 2023 Illegal Harms Consultation, p.25; Information Commissioner's Office (ICO) response to November 2023 Illegal Harms Consultation, p.19; Kooth Digital Health response to May 2024 Consultation on Protecting Children from Harms Online, p.14; LinkedIn response to November 2023 Illegal Harms Consultation, p.13; Mencap response to November 2023 Illegal Harms Consultation, p.12; National Trading Standards eCrime Team response to November 2023 Illegal Harms Consultation, p.11; NEXUS NI response to November 2023 Illegal Harms Consultation, p.14. We note that NEXUS NI made a similar point in May 2024 Consultation on Protecting Children from Harms Online, p.19; OnlyFans response to November 2023 Illegal Harms Consultation, p.7; Oxford Disinformation and Extremism Lab response to November 2023 Illegal Harms Consultation, p.15; Philippine Survivor Network response to November 2023 Illegal Harms Consultation, p.13; Pinterest response to May 2024 Consultation on Protecting Children from Harms Online, p.18; Refuge response to November 2023 Illegal Harms Consultation, p.18; Safe Space One Ltd response to November 2023 Illegal Harms Consultation, p.15; Segregated Payments LTD response to November 2023 Illegal Harms Consultation, p.11; Snap response to November 2023 Illegal Harms Consultation, p.16. We note that Snap made a similar point in May 2024 Consultation on Protecting Children from Harms Online, p.23; Vinted response to November 2023 Illegal Harms Consultation, p.13; WeProtect Global Alliance response to November 2023 Illegal Harms Consultation, p.19.

[1550] Are, C. response to November 2023 Consultation, p.13; Bereaved Families for Online Safety response to November 2023 Illegal Harms Consultation, p.2; Cifas response to November 2023 Illegal Harms Consultation, p.15; Global Partners Digital response to November 2023 Illegal Harms Consultation, pp.18-19; Humanists UK response to November 2023 Illegal Harms Consultation, p.15; WeProtect Global Alliance response to November 2023 Consultation, p.19.

[1551] Global Partners Digital response to November 2023 Consultation, p.19.

[1552] Are, C. response to November 2023 Consultation, p.13.

[1553] Bereaved Families for Online Safety response to November 2023 Consultation, p.2.

10.9 Respondents flagged that there are risks in providing too much detail in terms and statements. Several expressed concerns about the impact the level of detail will have on the clarity of terms and statements.[1554] Some raised concerns about the potential for detailed terms and statements to jeopardise the effectiveness of safety measures, and the risk of perpetrators using the information to circumvent such measures.[1555] We address these points in the 'Risks' section.

10.10 We understand from responses to our November 2023 Consultation that some service providers would like the flexibility to supply their terms and statements across multiple documents.[1556] This point is also addressed in the 'Risks' section.

10.11 Wikimedia Foundation expressed the view that the measure does not account for decentralised services.[1557] We address this in the 'Who this measure applies to' section.

## Our decision

10.12 We have decided to proceed with the measure as proposed in our November 2023 Consultation. This measure codifies what the Act requires service providers to include in their terms and statements. The concerns raised in the responses did not persuade us that this measure is disproportionate or ineffective.

10.13 The full text of the measure can be found in our Illegal Content Codes of Practice for U2U services and Illegal Content Codes of Practice for search services. This measure is part of our Illegal Content Codes of Practice on terrorism, child sexual exploitation and abuse ('CSEA'), and other duties and we refer to this as measure ICU G1 for U2U services and ICS G1 for search services.

## Our reasoning

### How this measure works

10.14 We recommend that all service providers include the following provisions in their terms and statements for each of their services (as applicable):

- (U2U and Search) provisions specifying how individuals are to be protected from illegal content;
  - (U2U only) addressing how the service provider will minimise the length of time for which any priority illegal content is present, with separate sections addressing

[1554] Meta response to November 2023 Illegal Harms Consultation, p.30. We note that Meta made a similar point in May 2024 Consultation on Protecting Children from Harms Online, p.29; Roblox response to May 2024 Consultation on Protecting Children from Harms Online, p.25; Snap response to November 2023 Consultation, p.16; UK Interactive Entertainment (Ukie) response to November 2023 Illegal Harms Consultation, p.24. We note that Ukie made a similar point in May 2024 Consultation on Protecting Children from Harms Online, p.48.

[1555] Google response to November 2023 Illegal Harms Consultation, p.56; LinkedIn response to November 2023 Consultation, p.13; Match Group response to November 2023 Illegal Harms Consultation, pp.13-14; Meta response to May 2024 Consultation, p.28; Mid Size Platform Group response to November 2023 Illegal Harms Consultation, p.10. We note that Mid Size Platform Group made a similar point in May 2024 Consultation on Protecting Children from Harms Online, p.11; Snap response to November 2023 Consultation, p.16; [✂].

[1556] Meta response to November 2023 Consultation, p.30. We note that Meta made a similar point in May 2024 Consultation, p.29; Ukie response to November 2023 Consultation p.24. We note that Ukie made a similar point in May 2024 Consultation, p.48.

[1557] Wikimedia Foundation response to November 2023 Illegal Harms Consultation, p.31.

terrorism content, child sexual exploitation and abuse (CSEA) content, and other priority illegal content;

> (U2U only) addressing how the service provider will act to remove illegal content when alerted by a member of the public to the presence of illegal content or made aware of it in any other way.

- (U2U and Search) provisions giving information about any proactive technology used for the purposes of compliance with the illegal content safety duties (including the kind of technology used, when it is used, and how it works).[1558]
- (U2U and Search) provisions specifying the policies and processes that govern the handling and resolution of relevant complaints.

10.15 Service providers must ensure that all the required provisions meet the clarity and accessibility standard required by the Act (see details in the section dedicated to our measure on 'Clarity and accessibility of terms and statements') regardless of the number of documents that constitute the provider's terms and statements, or where this information is located within their terms and statements.

## Benefits and effectiveness

10.16 This measure codifies provisions required by the Act and is important to ensure transparency around the steps service providers are taking to protect users from illegal content. This will empower users to make more informed choices about what services to use, thereby reducing the risk of online harm.

10.17 Some respondents from civil society organisations provided further suggestions around the types of information that U2U service providers should include in their terms and statements, or suggested that we should specify the information to be included in terms and statements.[1559]

10.18 While we acknowledge these points, we consider that excessive levels of detail in terms and statements about how systems and processes operate or are being used to protect users may require a disproportionate use of resources to keep up to date. This is particularly relevant where safety technology and processes are constantly evolving. Giving service providers the flexibility to manage the level of detail in terms and statements may help them ensure that their policies remain agile in changing circumstances. Granting service providers flexibility is also beneficial given the hugely diverse range of services in scope of the Act.

10.19 We did not receive any further evidence in our November 2023 or May 2024 Consultations to support the effectiveness of more prescriptive requirements. Therefore, we consider that this measure, as drafted, is sufficiently clear for providers to understand what is required of them and meet their relevant duties as set out in the Act.

[1558] For this purpose, 'illegal content safety duties' means the duties in section 10(2) and (3) of the Online Safety Act 2023 ('the Act') in relation to U2U service providers and the duties in section 27(2) and 27(3) of the Act in relation to search service providers.

[1559] Are, C. response to November 2023 Consultation, p.13; Bereaved Families for Online Safety response to November 2023 Consultation, p.2; Cifas response to November 2023 Consultation, p.15; Global Partners Digital response to November 2023 Consultation, pp.18-19; Humanists UK response to November 2023 Consultation, p.15; WeProtect Global Alliance response to November 2023 Consultation, p.19.

## Costs and risks

10.20 Service providers will need to ensure their terms and statements include provisions explaining how illegal content is actioned, how they use proactive technology to do this, and how they handle and resolve complaints (as set out in paragraph 10.14). We have not considered the costs of developing these provisions as this is a direct requirement of the Act.[1560]

10.21 Some respondents highlighted the risk of their terms and statements becoming extensive and hard to understand.[1561] To avoid lengthy documents, we consider that service providers can provide terms and statements in multiple documents. Furthermore, explanations or specifications about how a system or process works do not need to be repeated in a service providers' terms and statements, so long as it remains clear to users what information applies to each kind of priority illegal content and to each of the provisions, as set out in the 'How this measure works' section.

10.22 We recognise that service providers have concerns that the level of detail required in terms and statements may jeopardise the effectiveness of safety measures they have put in place, and that perpetrators may use the information to circumvent these safety measures.[1562] To mitigate this risk, some service providers may wish to manage the level of detail contained in their terms and statements, which is permissible as long as the documents still meet the requirements of the Act. In accordance with the Act, service providers must include provisions in their terms and statements specifying how users of their service are protected from encountering illegal content and our measures permit service providers flexibility in how they do so.

## Rights impact

10.23 The measure has no additional effect on rights, as it is a direct requirement of the Act.

## Who this measure applies to

10.24 This measure will apply to all U2U and search service providers, as the Act requires them to include in their terms and statements the relevant provisions mentioned in paragraph 10.14.

10.25 Service providers are likely to have different approaches to tackling the risk of illegal content appearing on their service. This will affect the provisions that they are required to include in their terms and statements. This measure allows for flexibility around the information service providers include in their terms and statements, as long as they meet the requirements in a way that is appropriate, accurate, and proportionate to their service.

[1560] Sections 10(5), 10 (6), 10(7), 21(3), 27(5), 27(6), 27(7) and 32(3) of the Act.

[1561] Meta response to November 2023 Consultation, p.30. We note that Meta made a similar point in May 2024 Consultation, p.29; Roblox response to May 2024 Consultation, p.25; Snap response to November 2023 Consultation, p.16; Ukie response to November 2023 Consultation, p.24. We note that Ukie made a similar point in May 2024 Consultation, p.48.

[1562] Google response to November 2023 Consultation, p.56; LinkedIn response to November 2023 Consultation, p.13; Match Group response to November 2023 Consultation, pp.13-14; Meta response to May 2024 Consultation, p.28; Mid Size Platform Group response to November 2023 Consultation, p.10. We note that Mid Size Platform Group made a similar point in May 2024 Consultation, p.11;  Snap response to November 2023 Consultation, p.16; [✂].

10.26 As noted in paragraph 10.11, Wikimedia Foundation expressed concern that decentralised services are not accounted for as part of our measure.[1563] To account for the range of service providers within the scope of our measures, we have chosen not to be prescriptive about how service providers should comply with this measure. Providers of decentralised services will need to ensure the provisions required by the Act are available in their terms and statements in a manner that is clear and accessible for all users. Beyond these requirements, service providers will be able to develop and design their terms and statements – including any additional provisions they may choose to include – in the way that best suits their service.

## Conclusion

10.27 For the reasons outlined in this section, we have decided to proceed with recommending the measure in the terms proposed in our November 2023 Consultation. All U2U and search service providers should include provisions in their terms and statements regarding the protection of individuals from illegal content, any proactive technology used, and information on how complaints are handled and resolved.

10.28 This measure is part of our Illegal Content Codes of Practice on terrorism, CSEA, and other duties and is referred to as ICU G1 for U2U services and ICS G1 for search services.

# Measure on additional requirements on terms and statements for Category 1 and Category 2A services

10.29 In our May 2024 Consultation, we proposed that providers of all Category 1 and Category 2A services summarise the findings of their most recent illegal content risk assessment in their terms or statement.

## Summary of stakeholder feedback[1564]

10.30 Many respondents were supportive of this measure and agreed that Category 1 and 2A service providers should summarise in their terms or statement the findings of their most recent illegal content risk assessment.[1565]

10.31 Microsoft expressed the view that Category 1 service providers should be given flexibility about where they choose to publish the summary of the illegal content risk assessment. They suggest that to avoid confusion for users and to ensure clarity of the terms and statements, the risk assessment and the terms and statements could be presented in

[1563] Wikimedia Foundation response to November 2023 Consultation, p.31.

[1564] Note this list is not exhaustive – further responses can be found in Annex 1.

[1565] Canadian Centre for Child Protection (C3P) response to May 2024 Consultation on Protecting Children from Harms Online, p.27; Centre for Excellence for Children's Care and Protection (CELCIS) response to May 2024 Consultation on Protecting Children from Harms Online, p.16; Children's Commissioner for England response to May 2024 Consultation on Protecting Children from Harms Online, p.68; Dean, J. response to May 2024 Consultation, p.18; Kooth Digital Health response to May 2024 Consultation, p.14; National Society for the Prevention of Cruelty to Children (NSPCC) response to May 2024 Consultation on Protecting Children from Harms Online, p.60; Scottish Government response to May 2024 Consultation on Protecting Children from Harms Online, p.17; Welsh Government response to May 2024 Consultation on Protecting Children from Harms Online, p.13.

separate locations.[1566] Pinterest made a similar point about clarity, and suggested flexibility in the format and location of where the risk assessment is published.[1567] We address this in the ‘How this measure works’ section.

10.32 Microsoft also asked that we clarify “that risk assessment summaries should avoid any detail that might facilitate or enable efforts to subvert child protection measures or illegal content risk mitigations”.[1568] Pinterest expressed a similar concern about the risk of perpetrators using the information to circumvent safety measures.[1569] We respond to this point under the ‘Risks’ section.

10.33 The National Society for the Prevention of Cruelty to Children (NSPCC) suggested that to ensure improved transparency, we should be more prescriptive about the types of information that should be included in the summaries of the illegal content risk assessment. It raised concern that without more detail, service providers may not comply with the measure as intended.[1570] This point is addressed in the ‘Risks’ section.

## Our decision

10.34 We have decided to proceed with the measure as proposed in our May 2024 Consultation. This measure codifies the requirement in the Act for Category 1 and Category 2A service providers to summarise in their terms or statement the findings of their most recent illegal content risk assessment. The concerns raised in the responses did not persuade us that this measure is disproportionate or ineffective.

10.35 The full draft of the measure can be found in our Illegal Content Codes of Practice for U2U services and Illegal Content Codes of Practice for search services. This measure is part of our Illegal Content Codes of Practice on terrorism, CSEA, and other duties and we refer to this as measure ICU G2 for U2U services and ICS G2 for search services.

## Our reasoning

### How this measure works

10.36 We recommend that Category 1 and 2A service providers summarise the findings of their most recent illegal content risk assessment in their terms and statements, including details about the levels of risk and the nature and severity of potential harm to individuals.[1571]

10.37 We recognise the ask for flexibility around where the summaries of the findings of the illegal content risk assessment are published. However, it is a direct requirement of the Act that the findings of the risk assessments are summarised in the terms or statements. The Act is not prescriptive about the number of documents that make up terms or statements. Our measure provides flexibility for service providers to choose to publish their terms and statements via multiple documents if this better suits their services, provided they meet the clarity and accessibility standard required by the Act (see details in the section dedicated to our measure on ‘Clarity and accessibility of terms and statements’).

[1566] Microsoft response to May 2024 Consultation on Protecting Children from Harms Online, p.15.

[1567] Pinterest response to May 2024 Consultation, p.18.

[1568] Microsoft response to May 2024 Consultation, p.15.

[1569] Pinterest response to May 2024 Consultation, p.18.

[1570] NSPCC response to May 2024 Consultation, pp.59-60.

[1571] See sections 10(9) and 27(9) of the Act for full details of these duties.

## Benefits and effectiveness

10.38 This measure codifies provisions required by the Act. It will help users better understand the risk of harm to individuals from illegal content in the context of their use of a service, and in the context of the measures, systems, and processes the service provider has in place to protect them. It will also lead to increased transparency around the risks of using a service and will empower users to make more informed choices. Greater transparency for users does not change the operation of the illegal harms safety duties; the onus remains on service providers to protect users from illegal content (rather than on users to protect themselves, even if they are empowered to do so).

10.39 This measure will help users make informed choices about whether and how they wish to use a service or allow children in their care to use it.

## Costs and risks

10.40 Service providers that do not currently include a summary of their most recent illegal content risk assessment in their terms and statements will need to do so. We have not considered the costs of including this information as this is a direct requirement of the Act for Category 1 and Category 2A service providers.

10.41 As discussed in paragraph 10.32, Microsoft and Pinterest highlighted the risk of perpetrators using the detail provided in the summaries of the findings of the illegal content risk assessment to undermine safety measures. We acknowledge this risk and are giving service providers the flexibility to comply with this measure in a way that does not equip perpetrators.

10.42 We also acknowledge the NSPCC's concern (paragraph 10.33) that this measure is not prescriptive as to the level of detail that service providers are required to include. However, we think it is important to allow some flexibility for service providers to ensure they are not providing information that could lead to unintended consequences, such as the risk of perpetrators learning how to circumvent protections. There is also a risk that in being more prescriptive, we may not impose the most suitable requirements. This is particularly the case given the diverse range of services in scope of the Act and the fast-moving nature of technological development. At this stage, we are only imposing the requirements as set out in the Act.

## Rights impact

10.43 The measure has no additional effect on rights, as it is a direct requirement of the Act.

## Who this measure applies to

10.44 This measure applies to all Category 1 and Category 2A service providers as required by the Act. Categorisation ensures that only service providers that reach a certain threshold of users and have high-risk functionalities are required to meet additional duties under the Act. Our proposed categorisation advice, submitted to the Secretary of State, gives a detailed overview of the role of categorisation and an outline of what will be required of categorised service providers.[1572]

[1572] Ofcom, 2024. Categorisation: Advice Submitted to the Secretary of State. [accessed 29 November 2024]

## Conclusion

10.45 This measure remains unchanged from the measure proposed in our May 2024 Consultation. Category 1 and 2A service providers should summarise the findings of their most recent illegal content risk assessment in their terms and statements. This measure will be contained in our Illegal Content Codes of Practice on terrorism, CSEA, and other duties and is referred to as measure ICU G2 for U2U services and ICS G2 for search services.

# Measure on clarity and accessibility of terms and statements

10.46 In our November 2023 Consultation, we recommended all U2U and search service providers should ensure that relevant provisions included in terms and statements regarding the protection of individuals from illegal content are clear and accessible. This measure is intended to secure compliance with the duties in the Act relating to the clarity and accessibility of the provisions set out in the measure on 'Substance of terms and statements'.[1573]

10.47 We found there to be four key areas for how these provisions can be deemed clear and accessible: findability, layout and formatting, language, and usability. We considered that an outcomes-based approach, which would set high-level expectations for services in these areas, would best accommodate the range of services in scope of regulation and allow services more flexibility in how they meet their duties.

## Changes to the measure in our May 2024 Consultation

10.48 In our November 2023 Consultation, we recommended that service providers write their terms and statements to a reading age comprehensible for the youngest individual permitted to agree to them. Subsequently, in the May 2024 Consultation, we consulted on updating this measure. We proposed altering the wording to read: "to a reading age comprehensible for the youngest individual permitted to use the service without consent from a parent or guardian".[1574]

## Summary of stakeholder feedback[1575]

10.49 Many stakeholders were supportive of this measure and agreed on the importance of clear and accessible terms.[1576] Several stakeholders stated the importance of accessibility of language.

[1573] Sections 10(8), 27(8), 21(3), 32(3) of the Act.

[1574] May 2024 Consultation on Protecting Children from Harms Online, p.288.

[1575] Note this list in not exhaustive – further responses can be found in Annex 1.

[1576] ACT – The App Association response to May 2024 Consultation on Protecting Children from Harms Online, p.23; Are, C. response to November 2023 Consultation, p.13; Betting and Gaming Council response to November 2023 Consultation, p.10; Big Brother Watch response to November 2023 Consultation on Protecting Children from Harms Online, p.10; British and Irish Law Education and Technology Association response to November 2023 Consultation, p.13; C3P response to November 2023 Consultation, p.23; Cats Protection response to November 2023 Illegal Harms Consultation, p.12; CELCIS response to May 2024 Consultation, p.16; CELE response to November 2023 Consultation, p.11; Children's Commissioner for England response to May

10.50 The Scottish Government expressed its view that policies should be drafted in a way that is clear and accessible for children in order to protect younger users.[1577]

10.51 [✁] argued that service providers "may need to consider widening the targeted age for information to ages other than the prescribed age limits of their site".[1578]

10.52 Other stakeholders also highlighted the need for terms and statements to be accessible to children, as well as users with disabilities or learning difficulties.[1579] [1580]

10.53 We address these points in the 'Benefits and effectiveness' section.

10.54 Furthermore, in response to our May 2024 Consultation, the Children's Commissioner for England suggested lowering the reading age terms and statements are written for.[1581] We also address this feedback in the 'Benefits and effectiveness' section.

10.55 Snap expressed the view that our estimated costs for achieving clarity and accessibility were too low.[1582] We address this concern in the 'Responses on costs' section.

10.56 Protection Group International said that even where terms of service are easily accessible, users can "still post, share, and distribute illegal content".[1583] We address this concern in the 'Benefits and effectiveness' section.

10.57 In response to our May 2024 Consultation, one civil society group [✁] questioned the likelihood of children engaging with services' terms and statements.[1584] Services also raised

2024 Consultation, pp.66; [✁]; Name withheld 5 response to November 2023 Consultation, p.13; Dwyer, D. response to November 2023 Consultation, p.7; Evri response to November 2023 Consultation, p.7; Federation of Small Businesses response to November 2023 Consultation, p.4. We note that Federation of Small Businesses made a similar point in May 2024 Consultation, p.7; 5Rights Foundation response to November 2023 Consultation, p.25; Kooth Digital Health response to May 2024 Consultation, p.14; Match Group response to November 2023 Consultation, p.13; Meta response to May 2024 Consultation, p.29; National Trading Standards eCrime Team response to November 2023 Consultation, p.11; NEXUS NI response to November 2023 Illegal Harms Consultation, p.14. We note that NEXUS NI made a similar point in May 2024 Consultation, p.19; OnlyFans response to November 2023 Consultation, p.7; Oxford Disinformation and Extremism Lab response to November 2023 Consultation, p.15; Philippine Survivor Network response to November 2023 Consultation, p.13; Pinterest response to May 2024 Consultation, pp.17-18; Refuge response to November 2023 Consultation, p.18; Safe Space One Ltd response to November 2023 Consultation, p.15; Scottish Government response to May 2024 Consultation, p.17; Segregated Payments LTD response to November 2023 Consultation, p.11; Snap response to November 2023 Consultation, p.17; Welsh Government response to November 2023 Illegal Harms Consultation, p.4. We note that Welsh Government made a similar point in May 2024 Consultation, p.13.

[1577] Scottish Government response to November 2023 Illegal Harms Consultation, p.9.

[1578] [✁].

[1579] CELCIS response to May 2024 Consultation, p.16; Children's Commissioner for England response to November 2023 Illegal Harms Consultation, p.22; 5Rights Foundation response to November 2023 Consultation, p.25; Internet Matters response to November 2023 Illegal Harms Consultation, p.17; Scottish Government response to May 2024 Consultation, p.17.

[1580] Children's Commissioner for England response to November 2023 Consultation, p.22; Glitch response to November 2023 Consultation, p.10; Mencap response to November 2023 Consultation, p.12; The Cyber Helpline response to November 2023 Illegal Harms Consultation, p.16.

[1581] Children's Commissioner for England response to May 2024 Consultation, pp.66-68.

[1582] Snap response to November 2023 Consultation, p.17.

[1583] Protection Group International response to November 2023 Illegal Harms Consultation, p.9.

[1584] [✁].

concerns as to the difficulty in designing terms and statements to be accessible to children.[1585] We also address these points in the 'Benefits and effectiveness' section.

10.58 Additionally, in response to our May 2024 Consultation, some stakeholders said that to make terms and statements clear and accessible to children and young people, terms and statements should be developed with input from them.[1586] This point is addressed in the 'Benefits and effectiveness' section.

## Our decision

10.59 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation, with the amendment proposed in our May 2024 Consultation. The measure now reads:

- All U2U and search service providers should ensure that relevant provisions included in terms and statements regarding the protection of individuals from illegal content, are clear and accessible to a reading age comprehensible for the youngest individual permitted to use the service without consent from a parent or guardian.

10.60 The full text of the measure can be found in our Illegal Content Codes of Practice for U2U services and Illegal Content Codes of Practice for search services. This measure is part of our Illegal Content Codes of Practice on terrorism, CSEA, and other duties and we refer to this as measure ICU G3 for U2U services and ICS G3 for search services.

## Our reasoning

### How this measure works

10.61 We recommend that service providers ensure the following four factors for clear and accessible terms and statements are taken into consideration when drafting provisions:

- Findability: Provisions should be easy to find, in that they are locatable within the terms or statement and are clearly signposted to the public (including to those who do not use or are not signed up for the service).
- Layout and formatting: Provisions should be laid out and formatted in a way that helps users read and understand them.
- Language: Provisions should be written to a reading age comprehensible for the youngest individual permitted to use the service without the consent of a parent or guardian.
- Usability: Provisions should be designed to be compatible with assistive technologies, including keyboard navigation and screen reading technology.

[1585] Skyscanner response to May 2024 Consultation on Protecting Children from Harms Online, p.17; Ukie response to May 2024 Consultation, p.48.

[1586] Scottish Government response to May 2024 Consultation, p.17; Yoti response to May 2024 Consultation on Protecting Children from Harms Online, p.36.

## Benefits and effectiveness

### Benefits

10.62 Clear and accessible terms and statements ensure users can find reliable and up-to-date information about the safety practices of regulated service providers.

10.63 As we mentioned in our November 2023 Consultation, clear presentation of provisions can help users find and understand relevant information. There are a range of techniques which have been shown to be effective at improving user understanding of terms and statements.

10.64 This is illustrated in research conducted by the Behavioural Insights Team ('BIT'), which examined different approaches to improving user understanding of contractual terms.[1587] For instance, BIT found that using icons in conjunction with a summary of key terms increased user comprehension scores by 34% compared to a control which just had a link to terms and conditions.[1588] Research carried out by the Danish Competition and Consumer Authority also found that icon summaries increased user comprehension scores by 38%.[1589] Colour ratio and contrast is also highlighted by the Web Content Accessibility Guidelines, which recommend a 4.5:1 ratio colour contrast between body text and background.[1590]

10.65 Additionally, some users with a disability may require certain tools to make use of the provisions. For example, users with visual or motor impairments may be dependent on using a keyboard to navigate apps and webpages, while screen readers make content on a screen accessible for those who are unable to see it.[1591] [1592] The Web Content Accessibility Guidelines encourage reading sequences to be programmatically determinable, which is important for those using assistive technologies, and keyboard accessible amongst others.[1593]

10.66 This measure specifies in relatively high-level terms what services should do to ensure their terms and statements are clear and accessible, rather than specifying in detail what they should do. We would expect approaches to accessibility to vary from service to service, subject to their service's features and design, and on that basis recommend providers are best placed to decide how to ensure information is accessible to disabled people. This approach is beneficial as it gives service providers a degree of flexibility regarding how they discharge their duties in this area. This is important given the diverse range of services in scope of the Act and the fast-moving nature of technology.

[1587] The Behavioural Insights Team, 2023. Who we are. [accessed 29 November 2024].

[1588] The Behavioural Insights Team, 2019. Best practice guide: Improving consumer understanding of contractual terms and privacy policies: evidence-based actions for businesses. p.12 [accessed 29 November 2024]. We note that BIT found that using icons with long blocks of text did not work very well. They compared a long privacy policy with no icons to an identical policy that was illustrated with over 20 icons but found that icons did not help customers understand the policy better in that case. This points to the importance of combining icons with short, easy to understand information.

[1589] Danish Competition and Consumer Authority, 2018. Improving the effectiveness of terms and conditions in online trade. *Competitive Markets and Consumer Welfare*, 15, p.5 [accessed 29 November 2024].

[1590] Web Accessibility Initiative, 2023. Understanding SC 1.4.3: Contrast (Minimum) (Level AA). [accessed 31 October 2024].

[1591] Web Aim, 2022. Keyboard accessibility. [accessed 29 November 2024].

[1592] Royal National Institute of Blind people, 2023. Screen reading software. [accessed 29 November 2024].

[1593] Web Accessibility Initiative, 2018. Web Content Accessibility Guidelines (WCAG) 2.1 W3C Recommendation 05 June 2018. [accessed 31 October 2024].

### Effectiveness

10.67 As detailed in paragraphs 10.50 to 10.52, several stakeholders mentioned that consideration should be given to ensuring that terms and statements are clear and accessible for children, and users who have disabilities or learning difficulties.[1594] [1595]

10.68 The feedback around giving more consideration to children was taken into account when we proposed an addition to our measure in the May 2024 Consultation. We believe the updated language better reflects the way that many U2U and search service providers require parental or guardian consent for children under a certain age to agree to terms and statements. This is because even when a service provider makes effort to draft terms and statements simply and clearly, children will often need support to understand the providers' public-facing information.

10.69 Service providers still have a duty to make available clear and accessible terms and statements that will empower children to have safer experiences online, both independently and with the adults who care for them.

10.70 We recognise the importance of accounting for adults and children with disabilities or learning difficulties. To achieve this, our four-factor approach, described in the 'How this measure works' section, sets out that service providers should consider the findability and usability of these provisions, as well as how they are laid out and formatted, and the language used to describe them. By implementing these four elements, we expect service providers to compile and present certain provisions within their terms and statements in a way that is clear and accessible to all users on their service, including those who may have disabilities or learning difficulties.

10.71 We acknowledge feedback that to make terms and statements accessible to children and young people, service providers should consult with children and young people when developing their terms and statements.[1596] Whilst we consider that this approach has benefits and service providers can choose to develop their terms and statements in this way, it is not an approach that is necessary in order for services to meet their duties under the Act. We consider that terms and statements can be made accessible to children by service providers following our outcomes-based four-factor approach, set out in the 'How this measure works' section.

10.72 We understand that not all children may engage with terms and statements. We are recommending that services' terms and statements are "written to a reading age comprehensible for the youngest individual permitted to use the service without consent from a parent or guardian". This is to encourage those children who wish to be informed about a service to understand terms and statements without the help of an adult, as well as

[1594] CELCIS response to May 2024 Consultation, p.16; Children's Commissioner for England response to November 2023 Consultation, p.22; [✂]; 5Rights Foundation response to November 2023 Consultation, p.25; Internet Matters response to November 2023 Consultation, p.2; Scottish Government response to November 2023 Consultation, p.9. We note that Scottish Government made a similar point in May 2024 Consultation, p.17.

[1595] Children's Commissioner for England response to November 2023 Consultation, p.22; Glitch response to November 2023 Consultation, p.10; Mencap response to November 2023 Consultation, p.12; The Cyber Helpline response to November 2023 Consultation, p.16.

[1596] Scottish Government response to May 2024 Consultation, p.17; Yoti response to May 2024 Consultation, p.36.

enable children who are unable to use a service without consent from a parent/guardian to fully understand terms and statements with the help of an adult.

10.73 Further, we recognise the limitations of clarity and accessibility measures in ensuring that children can understand a service providers' terms and statements. Such documents may by their nature be long and complex, despite a service provider's best efforts to simplify and streamline them. However, there is still a duty on service providers to make terms and statements as clear and accessible as is reasonably possible.

10.74 We acknowledge the validity of the Children's Commissioner for England's suggestion to lower the reading age to which terms and statements are drafted, to ensure that they are easily accessible to a wide range of users.[1597] However, there is a limit on the extent to which lowering reading age ensures all users can understand terms and statements, and to how much service providers can simplify the language of their terms and statements to account for the age of the youngest possible user (who may be well below the permitted age of use).[1598] Therefore, we did not consider this to be proportionate or appropriate for the variety of services in scope.

10.75 Further, we have not found a sufficient body of evidence to determine an alternative reading age that would better deliver clear and accessible terms and statements across the range of service providers within the measure's scope. That is why reading age is only one part of our four-factor recommendation for clear and accessible terms.

10.76 We acknowledge stakeholder feedback that even with clear and accessible terms and statements, there will always be some level of risk to users from illegal content on a service.[1599] We recognise that service providers cannot entirely prevent harm. However, the Act makes it the duty of service providers to assess and manage safety risks arising from content and conduct on their service.

## Costs

10.77 The costs associated with this measure do not directly vary with the number of users of a service. The costs will therefore tend to represent a higher share of revenue for providers of smaller services.

10.78 The costs will depend on the length of the relevant sections addressing the recommended provisions, which is likely to vary between service providers. More complex services will require longer provisions-related sections to comply with the measure, which will increase the cost. Costs are also likely to be higher for services that permit younger users to use the service without consent from a parent or guardian. This is because it is likely to be more challenging to make terms and statements comprehensible for younger users.[1600]

[1597] Children's Commissioner for England response to May 2024 Consultation, pp.66-68.

[1598] 5Rights Foundation response to November 2023 Consultation, p.25: "While we agree that the reading age of the terms should be understandable to the youngest person able to agree to them, we would note that this would not necessarily mean that a 13-year-old, for example, would understand the contract they were entering in to".

[1599] Protection Group International response to November 2023 Consultation, p.9.

[1600] We have recalculated the estimates since the November 2023 Consultation in line with the latest wage data released by the Office of National Statistics (ONS). However, since our cost estimates are rounded, the estimates have not changed when using the updated wage assumptions.

10.79 In the following paragraphs, we review the costs associated with each of our four factors recommended for clear and accessible terms or statements.

### Findability

10.80 Service providers that do not already have publicly available provisions are likely to incur a one-off design and engineering cost when making the necessary user interface changes to meet the measure's requirements. We estimate the one-off cost to be between £2000 and £5000 for most service providers (and potentially significantly less for smaller service providers).[1601] There may also be some smaller ongoing maintenance costs. We consider that the costs of allowing the terms and statements to be found are directly related to the requirement in the Act to make them accessible.

### Layout and formatting

10.81 Service providers may need to edit the formatting of the provisions to facilitate user understanding by adding icons, bullet points, subtitles, and white space. Service providers may also need to change the text format, size, and colour relative to the background so that the text is easy to read. The measure does not make specific recommendations in this area, allowing service providers the flexibility to decide how best to help users read and understand their terms or statements. The total cost will depend on the extent of revisions required by service providers and the specific choices made to achieve the required outcome. While these changes are likely to incur one-off costs, service providers will also need to ensure they maintain a suitable layout and formatting whenever they revise the provisions. We anticipate the costs of this will be similar to the considerations detailed in the 'Findability' section, with an overall one-off cost of between £2000 and £5000 (plus some smaller ongoing maintenance costs).

### Language

10.82 Service providers may incur the additional cost of reviewing provisions to ensure they are expressed in language likely to be comprehensible to the youngest individual permitted to agree to them. Providers may then need to revise the language used to ensure that it is compliant with this measure. As in our November 2023 Consultation, we do not think it proportionate to recommend that terms and statements be made available in specific languages. If a service operated exclusively in a non-English language, there would not be an expectation for these to be translated in English.

10.83 The total cost would depend on the extent to which the provisions need to be revised. Substantial changes may be required if the age of the youngest individual permitted to agree to the provisions is misaligned with the current reading age the provisions are written to. While making these changes is likely to be a one-off cost, providers need to ensure that they retain the same comprehensibility in language whenever they update provisions. We estimate that simplifying 800 words of text from a reading age of 16 to a reading age of 13

[1601] This assumes it would take up to five working days for a relevant employee to research the best ways to meet the requirements (assuming their salary is similar to a software engineer) and up to five working days for a software engineer to implement the changes. For many service providers, it may take less time to research and implement any changes. See Annex 5 for a detailed description of our salary assumptions.

would take a suitably qualified employee three days, costing the service provider between £500 and £1500.[1602]

### Usability

10.84 Service providers may need to make one-off design changes to ensure the relevant provisions are keyboard-navigable and compatible with screen reading tools. While these changes are likely to be minimal and low-cost, some providers may face higher costs (for example, if 'skip links' need to be added, or if the levels of headings used in provisions are incorrectly labelled). We anticipate the one-off costs to be similar to those for the 'Findability' section at between £2000 and £5000, with some smaller ongoing maintenance costs.

### Responses on costs

10.85 We received minimal feedback on our costs analysis for this measure.[1603] However, one service provider, [✂], argued that our estimated costs for achieving clarity and accessibility were far too low, though it still supported the guidance set out in this measure. [✂].[1604] We accept that for some services the costs could be higher than the ranges we have set out in paragraphs 10.80 to 10.84. This is especially likely to be the case if the relevant provisions are longer and more complex, if younger users are permitted to use the service without consent from a parent or guardian, and if the provider aims to achieve a very high standard in its terms of service. [✂].[1605] While service providers may choose to go beyond our recommendations, such as by translating the terms and statements, we recognise that this would increase costs.

10.86 While costs could be substantially higher for some service providers, we remain of the view that for the majority, the costs of applying these recommendations will be relatively small, particularly given that the requirements of the measure are framed in relatively high-level terms such that services have a significant degree of flexibility about how to implement it. Furthermore, services are required by the Act to ensure relevant provisions are clear and accessible. As such, most of the costs of this measure relate to these specific requirements in the Act, over which we have no discretion.

## Rights impact

10.87 We have carefully considered whether this measure would constitute an interference with users' or service providers' freedom of expression or association rights, or an infringement on users' privacy rights. Our conclusion is that this will not be the case.

10.88 We consider this proposed measure to be of benefit to users in that it will help them understand how a service provider protects them from content that might be harmful. It will also make them aware of the user empowerment tools and the reporting and complaints mechanisms available to them. These benefits will have positive effects on users' rights to privacy and to freedom of expression and association.

[1602] Assuming a salary similar to the 'professional occupations' category within the annual survey of hours and earnings (ASHE) data. See Annex 5 for a detailed description of our salary assumptions.

[1603] We received some feedback on the general cost assumptions (e.g. salary assumptions) that are fed into these costs. We consider that feedback in Annex 5.

[1604] [✂].

[1605] [✂].

### Who this measure applies to

10.89 This measure applies to all U2U and search service providers because the Act requires them to ensure that the provisions in their terms and statements (outlined in the measure on 'Substance of terms and statements') are clear and accessible.

10.90 We recommend service providers ensure relevant provisions are drafted following a four-factor model that considers findability, layout and formatting, language, and usability. While we have focused on these four factors as one way of making terms and statements clear and accessible, the measure is not intended to be prescriptive in this respect. It allows service providers flexibility in how the required clarity and accessibility is achieved.

## Conclusion

10.91 The Act requires that services make their terms and statements clear and accessible. Doing so will deliver benefits to users. Our Codes provide relatively high-level recommendations about how service providers can comply with this requirement. This affords them with a degree of flexibility about how they comply, which is beneficial given the range of services in scope of the Act and the fast-moving pace of technological development. Whilst our measures will entail some costs, these are relatively modest and largely flow from direct requirements of the Act rather than from choices Ofcom has made. We have therefore decided to confirm this measure as proposed in the November 2023 Consultation and updated in the May 2024 Consultation.

10.92 This measure therefore remains unchanged from the updated measure proposed in our May 2024 Consultation. All U2U and search service providers should ensure that relevant provisions included in terms and statements regarding the protection of individuals from illegal content are clear and accessible.

10.93 This measure will be contained in our Illegal Content Codes of Practice on terrorism, CSEA, and other duties and is referred to as ICU G3 for U2U services and ICS G3 for search services.

# 11. User Access

## What is this chapter about?

Removing users and accounts that post the most harmful types of illegal content is an effective way of combatting the spread of such content. However, such restrictions need to be considered very carefully given the impact they have on freedom of expression. This chapter sets out the measure we are recommending in relation to the removal of user accounts operated on behalf of proscribed terrorist organisations, why we are recommending it, and to which user-to-user (U2U) services it should apply.

## What decisions have we made?

We are recommending the following measure:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| **ICU H1** | Providers should **remove a user account** from a service if there are reasonable grounds to infer it is **operated by or on behalf** of a **terrorist organisation proscribed by the UK Government.** | Providers of U2U services |

## Why have we made these decisions?

It very likely that the content generated, shared, or uploaded via accounts operated on behalf of proscribed organisation will amount to an offence. Removing such accounts should make it more difficult for these organisations to communicate and cause harm online. This will help protect users and society at large from the harm caused by online terrorist content.

## Introduction

11.1 User access concerns a user's entry on to a service and their ability to use the functionalities present on that service. This includes control of access throughout the user journey, including measures taken by service providers in response to identified illegal behaviour. Restrictions on user access are a potential means of reducing harm as they can constrain perpetrators from using a service and act as a deterrent against engaging in illegal conduct online.

11.2 User access measures are related to service providers' content moderation processes (such as detecting illegal content via automated or human moderation) and can be used as sanctions in response to upheld complaints. Terms of service play an important role in ensuring that users understand how their access to the service may be limited. These processes also ensure that users have appropriate information concerning potential redress where they believe the terms of service have been incorrectly applied.

11.3 We view a measure concerning user access as being related exclusively to user-to-user ('U2U') services and not search services. This is because users (1) are not usually required to hold accounts to use search services, and (2) do not use search services to upload or share content in the same manner as U2U services.

11.4 Under their illegal content safety duties in the Online Safety Act 2023 ('the Act'), regulated U2U service providers must take certain steps to reduce the risk of harm posed by illegal content to users of the service, as listed in section 10(2). The requirements in these sections include, where proportionate, "policies on user access to the service or particular content present on the service, including blocking users from accessing the service or particular content" (section 10(4)(d)).[1606]

11.5 In our November 2023 Illegal Harms Consultation ('November 2023 Consultation'), we considered recommending several different applications of a strikes and blocking system.[1607] We considered the following options for Illegal Content Codes of Practice ('Codes') measures:

- U2U services should employ a strikes and blocking system against users where they are found to have posted or shared illegal content or committed or facilitated illegal behaviour;
- U2U services should block users where they are found to have shared content relating to or facilitating certain offences where there is a risk of repeat behaviour. Specifically, they should:
  i) Block users where they are found to have shared content relating to or facilitating child sexual abuse material (CSAM); or
  ii) Remove accounts operated by or on behalf of proscribed organisations.

11.6 Having considered these options, we ultimately proposed the last option around the removal of accounts operation by or on behalf of proscribed organisations. We said we would do further work to develop a proposed codes measure on banning accounts that have shared CSAM. In Spring 2025 we will consult on a measure in this space.

## Structure of this chapter

11.7 This chapter will begin with a discussion of our proposed measure to recommend services remove accounts operated by proscribed organisations, including the stakeholder feedback on the proposed measure and our final decision.

11.8 We will then, in turn, review the feedback on the measures we did not propose to recommend, including our proposal not to recommend an identity verification measure.

# Measure on removing proscribed organisation accounts

11.9 In the November 2023 Consultation, we proposed that all U2U service providers should remove a user account from the service if they have reasonable grounds to infer it is operated by or on behalf of a terrorist group or organisation proscribed by the UK Government (a 'proscribed organisation'). Service providers can find a list of proscribed

[1606] Section 10(4)(d) of the Online Safety Act 2023.

[1607] A note on terminology: we have reflected that 'banning' is a more appropriate term to use in this context. As such, throughout this chapter, we have used the term 'banning' instead of 'block' or 'blocking', where appropriate. We only refer to 'block' when referring to what was proposed in our November 2023 Consultation. This avoids any potential confusion with the feature or function available to users who wish to block another user. We also note that the Act refers to 'banning' users (see e.g. sections 17(8), 18(12), and 71(3)(b)).

organisations linked in our Illegal Content Judgements Guidance ('ICJG') or on the UK Government website.[1608]

11.10 We considered this measure to be proportionate due to its likely effectiveness in reducing the amount of content amounting to a proscribed organisation offence ('proscribed organisation content') or other terrorism offence on services.

## Summary of stakeholder feedback[1609]

11.11 A range of stakeholders, including providers of regulated services and civil society organisations, expressed broad support for our proposed measure.[1610] A number of respondents agreed that service providers should remove accounts operated by or on behalf of proscribed organisations.[1611] Several also explained their view that the measure is proportionate to the likelihood of harm caused by such accounts and the content spread via those accounts.[1612] Additionally, some service providers indicated they already ban user accounts that spread illegal content.[1613] There were no significant concerns about the measure's technical feasibility or new evidence to suggest the measure is disproportionate.

11.12 In addition to support for this measure, stakeholders highlighted several areas for further consideration:

- Clarity on how to identify accounts.
- Proactive detection.
- Use of UK proscribed organisations list.
- Risk of removed users continuing to cause harm.
- Rights impacts.
- Private communications.
- Proportionality of our measure.

[1608] ICJG: chapter 2 'Terrorism'

[1609] Note this list in not exhaustive, and further responses can be found in Annex 1.

[1610] Centre for Competition Policy Illegal Harms response to November 2023 Illegal Harms Consultation, p.17; Name withheld 5 response to November 2023 Illegal Harms Consultation, p.13; Dwyer, D. response to November 2023 Illegal Harms Consultation, p.10; Federation of Small Businesses response to November 2023 Illegal Harms Consultation, p.4; INVIVIA response to November 2023 Illegal Harms Consultation, p.25; Local Government Association response to November 2023 Illegal Harms Consultation, p.14; Match Group response to November 2023 Illegal Harms Consultation, p.18; Mencap response to November 2023 Illegal Harms Consultation, p.15;]; Name withheld 4 response to November 2023 Illegal Harms Consultation, p.9; Nexus response to November 2023 Illegal Harms Consultation, p.19; Segregated Payments Ltd response to November 2023 Illegal Harms Consultation, p.13; Snap response to November 2023 Illegal Harms Consultation, p.24; South East Fermanagh Foundation (SEFF) response to November 2023 Illegal Harms Consultation, p.18; Tech Against Terrorism response to November 2023 Illegal Harms Consultation, p.9; The Cyber Helpline response to November 2023 Illegal Harms Consultation, p.19.

[1611] Federation of Small Businesses response to November 2023 Consultation, p.4; Match Group response to November 2023 Consultation, p.18; Meta and WhatsApp response to November 2023 Consultation, annex, pp.16-17; Snap response to November 2023 Consultation, p.24.

[1612] Centre for Competition Policy response to November 2023 Consultation, p.17; Tech Against Terrorism response to November 2023 Consultation, p.9.

[1613] Evri response to November 2023 Illegal Harms Consultation, p.9. See CSAM blocking measure and Annex 1 for others.

11.13 We summarise these themes in the following sections.

## Clarity on how to identify accounts

11.14 A number of stakeholders provided feedback regarding the threshold we recommended service providers use to determine if an account is run by or on behalf of a proscribed organisation. Some stakeholders said the 'reasonable grounds' threshold for this measure is ambiguous, and said it would make it challenging for service providers to determine whether an account meets the threshold.[1614] There were also concerns that the reasonable grounds threshold is insufficiently demanding and could lead to inconsistent content moderation decisions across services, reducing the effectiveness of the measure.[1615] Lastly, a stakeholder indicated that the threshold may establish a lower standard of evidence than required under UK criminal law.[1616] We also received feedback saying that service providers may find it difficult to determine whether an account is run by or on behalf of a proscribed organisation.[1617] We address these concerns in the 'How this measure works' section below.

## Proactive detection

11.15 One stakeholder expressed concern that our measure was recommending the proactive detection of illegal content, stating that it would be extremely cumbersome and disproportionate to implement for services that are not at high risk.[1618] We address this concern in the 'How this measure works' section below.

## Use of UK proscribed organisations list

11.16 Some stakeholders were apprehensive about the measure's use of the UK proscribed organisations list. We received feedback from stakeholders that expressed a preference for using other lists, such as those of the US and the United Nations Security Council.[1619] Another stakeholder raised concern that government lists could be weaponised by governments around the world against political opponents or other groups.[1620]

11.17 The Christchurch Call Advisory Network (CCAN) argued that terrorist organisations are sometimes a part of, or carry out, government functions. It said this measure would impede these functions and expressed a more general concern about the measure's reliance on a list of proscribed terrorist groups, rather than the context of the content posted on an account.[1621] We address these issues in the 'How this measure works' and 'Rights' sections below.

[1614] British and Irish Law, Education, and Technology Association (BILETA) response to November 2023 Illegal Harms Consultation, p. 17; Christchurch Call Advisory Network (CCAN) response to November 2023 Illegal Harms Consultation, pp.2-3; Electronic Frontier Foundation (EFF) response to November 2023 Illegal Harms Consultation, p.19; Google (confidential) response to November 2023 Illegal Harms Consultation, pp.72-73; Tech Against Terrorism response to November 2023 Consultation, pp.9-10.

[1615] Google response to November 2023 Illegal Harms Consultation, pp.3, 67-68.

[1616] CCAN response to November 2023 Consultation, p.2.

[1617] Tech Against Terrorism response to November 2023 Consultation, p.9.

[1618] Pinterest response to November 2023 Illegal Harms Consultation, p.11.

[1619] Mega response to November 2023 Illegal Harms Consultation, p.6.

[1620] Meta response to November 2023 Consultation, confidential annex, p.16.

[1621] CCAN response to November 2023 Consultation, pp.2-3.

## Risk of removed users continuing to cause harm

11.18 Several stakeholders expressed concern that the measure does not prevent users from creating new accounts to continue causing harm.[1622] We address this point in the 'Benefits and effectiveness' section below.

## Rights impacts

11.19 Some stakeholders highlighted that the measure's wording and use of proscribed organisation lists could lead to over-enforcement and erroneous account removal, potentially infringing on the right of users to freedom of expression.[1623] Other stakeholders emphasised the importance of human involvement in ensuring the accuracy of automated processes, and the need for robust mechanisms allowing users to appeal content or account removals.[1624] To address this, stakeholders recommended that service providers should (1) consider the nature of proscribed organisation content when assessing whether an account is operated by or on behalf of a proscribed organisation, rather than determining this solely on the volume of such content;[1625] (2) use human review, in addition to automated moderation tooling, for content moderation decisions to protect against wrongful and erroneous moderation practices;[1626] and (3) participate in collaborative efforts to identify trends and specific sources of risk to support moderation of their services, such as radicalisation pathways.[1627] We address this issue in the 'Rights impact' section below.

11.20 The CCAN also highlighted that, in practice, providers rarely attempt to contextualise illegal content and prefer not to allow any kind of proscribed organisation to have an account due to potential liability considerations (a practice called 'collateral censorship').[1628] We also note a relevant response by Oxford Disinformation and Extremism Lab (OxDEL) regarding our guidance to providers on how to identify illegal content, including terrorist content. OxDEL expressed concern that the scope of the approach to identifying illegal content could lead to overreach or abuse targeted at peaceful dissenters, civil society and academic researchers.[1629] We respond to this in the ICJG: chapter 2 'Terrorism', but also address it below in the 'Rights impact' section to the extent it is relevant to this measure.[1630]

## Private communications

11.21 [✂] raised concerns around the "difficulty of assessing context with regards to potentially illegal content…such context is often not readily available to services like ours, where content is located within a private space without the additional contextual information provided by features such as personal profiles, comments, or reposts. This lack

[1622] Institute for Strategic Dialogue (ISD) response to November 2023 Illegal Harms Consultation, p.12; SEFF response to November 2023 Consultation, p.18.

[1623] CCAN response to November 2023 Consultation, p.2; Centre for Competition Policy response to November 2023 Consultation, pp.3, 16; ISD response to November 2023 Consultation, p.12.

[1624] Federation of Small Businesses response to November 2023 Consultation, p.4; Microsoft response to November 2023 Consultation, p.20.

[1625] [✂]

[1626] Federation of Small Businesses response to November 2023 Consultation, p.4.

[1627] Centre for Competition Policy response to November 2023 Consultation, p.16.

[1628] CCAN response to November 2023 Consultation, p.3.

[1629] Oxford Disinformation and Extremism Lab (OxDEL) response to November 2023 Illegal Harms Consultation, p.8.

[1630] As explained in the ICJG: chapter 2 'Terrorism'.

of context makes it extremely difficult to assess the risk without jeopardising users' privacy and freedom of expression…".[1631]

11.22 WhatsApp also argued that an approach focussed on volume of content posted would not be actionable or appropriate for encrypted services.[1632]

11.23 We respond to these concerns in the 'How this measure works' section below.

### Proportionality

11.24 We also received general feedback regarding the proportionality of our total package of measures applicable to small services.[1633] We discuss this in the context of this measure in the 'Who this measure applies to' section.

## Our decision

11.25 We have decided to proceed with the measure broadly as proposed in our November 2023 Consultation. The full text of the measure can be found in our Illegal Content Codes of Practice for U2U services and is referred to as ICU H1. This measure is part of our Illegal Content Codes of Practice on terrorism.

## Our reasoning

### How this measure works

#### Terrorism offences and proscribed organisations

11.26 The measure only applies to user accounts operated by or on behalf of proscribed organisations. Proscribed organisations are those that have been banned by the UK Home Secretary following assessment against several factors set out in legislation, including the specific threat they pose to the UK.[1634] [1635]

11.27 Accounts operated by or on behalf of these organisations differ from other accounts in that any content generated, shared, or uploaded via that account is very likely to amount to an offence in the UK. Any such content is therefore likely to be priority illegal content, and even the setting up of an account is likely to amount to one or more priority offences. This is because, in addition to the priority terrorism offences relating to proscribed organisations, the priority offence of preparation of terrorist acts includes any conduct in

[1631] Name withheld 5 response to November 2023 Consultation, p.13.

[1632] WhatsApp response to November 2023 Consultation, annex, p.17.

[1633] The Global Network Initiative response to November 2023 Illegal Harms Consultation, p.8. See also chapter 'Our approach to developing Codes measures'.

[1634] Home Office, 2021. List of proscribed terrorist groups and organisations. [accessed October 29, 2024]

[1635] The UK government publishes a list of proscribed terrorist organisations. To proscribe an organisation, the Home Secretary must have a reasonable belief that the organisation is currently involved in terrorism, and that it is proportionate to proscribe it. The Home Secretary will make this decision having considered all relevant factors, including the specific threat a group poses to the UK. Proscription decisions require approval from both Houses of Parliament.

preparation for an act of terrorism.[1636] [1637] An act of terrorism includes any action intentionally taken for the benefit of a proscribed organisation.[1638]

11.28 Removing proscribed organisations' accounts should therefore protect users by preventing the service from being used for the commission of an offence, and by making it more difficult for these organisations to share illegal content.

11.29 As mentioned in paragraph 11.16, some stakeholders expressed concerns regarding the sole use of the proscribed organisations list.[1639] This measure will assist service providers to comply with their duties relating to illegal content as outlined in the Act. Illegal content includes, as explained, content which amounts to a proscribed organisation offence or other terrorism offence under UK domestic law. We therefore consider that, in order for the measure to assist providers to comply with their duties, it has to refer to organisations which are proscribed in the UK. The most effective way to achieve this is to refer to the list of UK proscribed organisations.

11.30 If service providers choose to refer to a different list of proscribed organisations, they will need to ensure their approach nevertheless catches all UK proscribed organisations in order to comply with this measure.

### Identifying accounts

11.31 In many cases, service providers may become aware of an account linked to one of these groups due to a piece of illegal content flagged through their moderation process. Service providers may also be made aware of an account by law enforcement, another service user, or a member of the public who identifies the content and requests it to be taken down via reporting or complaints processes.

11.32 A service provider should consider whether an account might be operated by or on behalf of a proscribed organisation when the service provider has:[1640]

- determined that content on the account amounts to a proscribed organisation offence;
- determined that content on the account is in breach of an equivalent standard as set out in the terms of service; or
- otherwise become aware that an account may be operated by or on behalf of a proscribed group (including as a result of a report or complaint).

[1636] The offences relating to proscribed organisations are as follows: belonging or professing to belong to a proscribed organisation; inviting support for a proscribed organisation; expressing an opinion or belief supportive of a proscribed organisation; arranging, managing or assisting in arranging or managing a meeting which the suspect knows to support or further the activities of a proscribed organisation or to be addressed by a person belonging or professing to belong to a proscribed organisation; addressing a meeting where the purpose of the address is to encourage support for a proscribed organisation or to further its activities; wearing an item of clothing or wearing, carrying or displaying an article in a public place in such a way or in such circumstances as to arouse reasonable suspicion that they are a member or a supporter of a proscribed organisation; publishing an image of any article in such a way or in such circumstances as to give rise to reasonable suspicion of membership or being a supporter of a proscribed organisation. Part 2 of the Terrorism Act 2000.

[1637] Section 5 of the Terrorism Act 2006; see also section 20(2) for interpretation.

[1638] Section 20(2) of the Terrorism Act 2006; and section 1(5) of the Terrorism Act 2000.

[1639] CCAN response to November 2023 Consultation, pp.2-3; Mega response to November 2023 Consultation, p.6; Meta response to November 2023 Consultation, confidential annex, p.16.

[1640] As explained in the ICJG: chapter 2 'Terrorism'.

11.33 As mentioned in paragraph 11.15, in response to our November 2023 Consultation, one stakeholder expressed concerns that our measure was recommending the proactive detection of content.[1641] For the avoidance of doubt, the measure does not stipulate that service providers should use proactive detection of illegal content. Nevertheless, if a service provider of its own volition chooses to take proactive steps to detect potentially illegal content then it may use the output of this proactive detection to inform its judgements as to whether a particular account is operated by or on behalf of a proscribed organisation.

11.34 We also recognise that service providers may not be certain that an account is operated by or on behalf of a proscribed organisation.[1642] We therefore expect service providers to remove an account when they have reasonable grounds to infer this. We consider this to be an appropriate threshold as it is consistent with the threshold for making an illegal content judgement under the Act. There are several factors that we consider may give rise to reasonable grounds for inference. These include a combination of user profile factors, though this list is not exhaustive.

- **Username** – The username may contain, refer to, or be that of a proscribed organisation or a known or listed alias for a proscribed organisation.
- **User profile images such as profile, account, or background images** – The user profile image may contain logos or symbols connected in some way to the proscribed organisation or the name of the group. This may include images which have been edited or otherwise obscured to evade detection by automated systems.
- **User profile information** – Other information fields attached to the account may suggest it is operated by or on behalf of a proscribed organisation. This may include the use of the organisation name in a user profile, bio, or in another descriptive field such as those describing a user's education, workplace, or political beliefs

11.35 These factors may not always be present when an account is operated by or on behalf of a proscribed organisation. As such, we consider that reasonable grounds to infer may also arise where a significant proportion of a reasonably sized sample of content recently posted on or via the user account amounts to a proscribed organisation offence. We do not consider it practicable to specify precisely how much content a service provider should consider for this purpose, as we expect that this would vary across both service providers and specific cases. If a provider chooses to review a sample of content to determine whether to remove an account, we expect this to happen only after it has become aware of a piece of proscribed organisation content.

11.36 As mentioned in paragraph 11.14, in response to the November 2023 Consultation, several stakeholders raised concerns about the 'reasonable grounds' threshold, including, but not limited to, the threshold being ambiguous, being insufficiently demanding, and leading to inconsistent decisions across different services.[1643] We acknowledge the concerns raised by stakeholders. However, as noted above, the 'reasonable grounds' threshold is consistent with the threshold for a service provider to make an illegal content judgement. This ensures

[1641] Pinterest response to November 2023 Consultation, p.11.

[1642] Tech Against Terrorism response to November 2023 Consultation, p.9.

[1643] BILETA response to November 2023 Consultation, p.19; CCAN response to November 2023 Consultation, pp.2-3; EFF response to November 2023 Consultation, p.19; Google response to November 2023 Consultation, p.3, 67-68; Tech Against Terrorism response to November 2023 Consultation, p.10.

consistency and operational ease when making a series of such judgements and deciding whether to remove an account in accordance with this measure.

11.37 Furthermore, as this measure will apply to all services, we believe the threshold also grants sufficient flexibility for providers to implement the measure in a way that is suitable for their service. A more prescriptive standard could be more difficult to apply to such a broad range of services and to the different ways a proscribed organisation account might present itself. For these reasons, we have decided not to change the threshold for this measure.

11.38 We recognise that it would be desirable, if it were possible, to provide very precise guidance on exactly when it is and is not appropriate for a service provider to infer that an account is run for or on behalf of a proscribed organisation. However, we do not think this is currently possible. These are judgments which can only be made based on the evidence available, which will vary greatly from case to case and will usually be incomplete. To have no measure on proscribed organisations would leave users exposed to harm.

11.39 We have therefore concluded that the proposed threshold and the factors we expect a provider to consider when making a decision under this measure are appropriate and effective, striking a reasonable balance between protecting users from the harm caused by proscribed organisations, respecting users' privacy and protecting them from incorrect action against their accounts.

### Privately communicated content

11.40 For the purpose of this measure, 'content' does not include content that has been privately communicated unless the relevant service provider has explicit consent to view the content in question (for example, having received a report about a private communication).

11.41 As mentioned in paragraph 11.21, [✂] raised concerns about the difficulty of assessing the context of content where it is held in private spaces and there are no other contextual factors to consider.[1644] We recognise this challenge and have designed the measure to only capture publicly communicated content. This is because a service provider viewing privately communicated content without explicit consent has implications for a user's right to privacy.[1645]

11.42 This measure applies to end-to-end encrypted services. However, we recognise that these services may face challenges in determining whether an account belongs to a proscribed organisation. We expect these services to determine whether an account is linked to a proscribed organisation via the alternative factors listed above, including (1) the user profile factors outlined in paragraph 11.34 or in cases where these factors are not present (2) a significant proportion of a reasonably sized sample of content recently posted on or via the user account amounts to a proscribed organisation offence, as outlined in paragraph 11.35.

11.43 It is also because we recognise that on some services, including some end-to-end encrypted services, the vast majority (if not all) of content uploaded, generated, or shared by users may be done privately. WhatsApp also argued that an approach focussed on volume of content posted would not be actionable or appropriate for encrypted services.[1646] However, on such services, what amounts to a 'reasonably sized sample' of content will depend on

[1644] Name withheld 5 response to November 2023 Consultation, p.13.

[1645] We discuss the privacy implications of this measure in more detail in the impact on users' rights section below.

[1646] WhatsApp response to November 2023 Consultation, annex, p.17.

the amount of content that the provider has explicit consent (and in the case of end-to-end encrypted services, is able) to view. For example, an end-to-end encrypted service provider may only be able to view content that other users have reported and may decide that that content represents a reasonably sized sample in the circumstances.

## Benefits and effectiveness

11.44 Effective user access measures can help service providers prevent illegal content from appearing and spreading on their services. They can also reduce the risk of repeat offending where removing an individual piece of content does not on its own sufficiently reduce risk, due to offending users continuing to carry out illegal activity while they retain access to the service.[1647] Preventing proscribed organisations from operating accounts on U2U services therefore has the potential to disrupt their activities and reduce their ability to disseminate illegal content which can pose a significant risk to UK users. The benefits of this measure are potentially significant given the harm proscribed organisations' activities can cause.

11.45 As stated in paragraph 11.27, proscribed organisations differ from all other users communicating illegal content in that any activity carried out on an account operated by or on behalf of a proscribed organisation is very likely to amount to an offence, and (to the extent it produces content) be priority illegal content. As providers are under a duty to remove illegal content from their service, we expect that the identification of a proscribed organisation account will also lead to the removal of much, if not all, of the content posted on that account on the basis that it is terrorist content.[1648] As such, we have determined this measure would be effective in preventing this type of illegal content from being spread on the service; the proscribed organisation is not only unable to disseminate further illegal content via the account in question, but also any illegal content posted prior to removal will no longer be accessible by other users.

11.46 We recognise that this measure does not prevent users from returning to a service after their account is removed. Some stakeholders felt that this could reduce the effectiveness of the measure, as a proscribed organisation could simply create a new account to continue causing harm.[1649] We recognise that preventing users from returning to a service is an important consideration for effectiveness. However, the prevention of removed users creating new accounts is a complex issue.

11.47 While we acknowledge the ability for proscribed organisations to create new accounts is a limitation of this measure, our view is that there is value in its recommendation because any disruption to the online activities of proscribed organisations is beneficial. Removing a user account reduces that user's ability to communicate with followers in the period following their removal. The creation of a new account adds further disruption, requiring more time and effort to rebuild networks with other users. We consider that the disruption caused to a proscribed group's online network and reach – and therefore the spread of

[1647] Although at the moment we are not currently recommending specific measures against these kinds of illegal harms, our Register of Risks ('Register') highlights the risk of repeated illegal behaviour for many of the kinds of illegal harm assessed, including but not limited to terrorism, hate, harassment, stalking, threats and abuse, drugs and weapons offences. In our view, removing users that participate in such activities from a service can be an effective way to reduce the prevalence of these online harms. See Register for discussion of each kind of illegal harm, and a focus on "User identification", "User networking" and "User communication" functionalities as particularly relevant to repeat offending.

[1648] Section 10(3)(b) of the Act.

[1649] ISD response to November 2023 Consultation, p.12; SEFF response to November 2023 Consultation, p.18.

illegal content – should overall reduce the risk of harm to users. We are though still developing evidence on the most appropriate and effective methods to prevent users from returning to a service.[1650]

11.48 As explained in chapter 2 of this Volume: 'Content moderation', we recognise that some service providers are unable to take down content. This means that, though the account has been removed, any illegal content disseminated by that account may still be accessible by other users. Where this is the case, we believe this measure will nevertheless reduce the risk of illegal harm by disrupting the proscribed organisation's network for the reasons explained above.

### Costs

11.49 We consider it unlikely that most service providers (including the large majority of smaller providers) are targets for proscribed organisations. Such providers will only incur the costs of assessing a suspicious account if they believe a piece of content amounts to a proscribed organisation offence (or is in breach of an equivalent standard in their terms of service), or if such an account is found, reported, or otherwise brought to their attention. If this does not occur, they will not incur any one-off or ongoing costs related to this measure, other than the costs of reading and understanding it.

11.50 Some service providers, such as larger social media providers, are more likely to be targeted by proscribed organisations and will incur greater costs for moderating their services. For such providers, the measure may involve the following costs.

- **Designing a process for staff to follow and providing associated training** – Service providers may consider it appropriate to set out a process for staff to follow when assessing whether an account is operated by or on behalf of a proscribed organisation. Developing this process is likely to require input from regulatory and/or legal staff, and the related costs are likely to vary depending on the size and type of service provider. Once the process is established, service providers will need to offer appropriate materials and training to enable staff to recognise a proscribed organisation account and confirm that it should be removed from the service. As section 10(3)(b) of the Act requires service providers to swiftly take down any illegal content as they become aware of it, it is likely that many service providers who have such content will need to train their staff to recognise illegal content. Adding a further step to identify a proscribed group is unlikely to incur significant additional costs.[1651]
- **Assessing suspicious accounts and removing as necessary** – The process of account reviews and removals will require content moderators to assess whether an account is operated by or on behalf of a proscribed organisation. Account removals may require further input from second-level support staff. In most cases, we do not expect account reviews and removals to be particularly complex or require technical expertise. Therefore, we do not expect them to incur significant costs. We recognise that service providers may use various methods to remove accounts. For example, a service

[1650] The following stakeholders submitted evidence including suggestions for best practices, to impede users (who had accounts blocked previously) from creating new accounts, which will be considered for future iterations of the Codes: [✂]; ISD response to November 2023 Consultation, p.12; [✂]; [✂]; SEFF response to November 2023 Consultation, p.18.

[1651] The additional step may be covered by adding extra hours to a one-off training session, which we estimate would add less than £200 per trained employee to the initial cost of attending training. Service providers service may also wish to provide additional training for content moderators on an ongoing basis.

provider may choose to automate the account removal process. This is likely to incur higher upfront costs but may result in lower ongoing costs. Working on the assumption that reviewing and removing an account takes two hours of a content moderator's time and one hour of a software engineer's time, this would lead to a cost per account reviewed and/or removed of approximately £140.[1652] We consider this estimate to be on the high side and, in practice, expect costs to be lower in most cases.[1653] The ongoing costs of this measure will depend on how frequently proscribed organisation content is shared on a service, as frequent posting may lead to faster detection or increased user complaints that prompt a review of an account.

- **Any costs incurred if a user were to appeal a decision to take down an account** – Establishing an appeal process for users who have been removed from service as a result of generating, uploading, or sharing illegal content is a requirement of the Act. However, this measure could result in more appeals, as a user whose account has been removed may complain. The provider will then incur costs in reviewing the appeal and restoring the account if the appeal is legitimate

## Rights impact

11.51 Because this measure recommends the removal of user accounts, it has important considerations and implications for freedom of expression, freedom of association, and privacy rights.

### Freedom of expression and freedom of association

11.52 s explained in 'Introduction, our duties, and navigating the Statement', as well as chapter 14 of this Volume: 'Statutory tests', Article 10 of the ECHR sets out the right to freedom of expression, which encompasses the right to hold opinions and to receive and impart information and ideas without unnecessary interference by a public authority. Article 11 of the ECHR sets out the right to associate with others. The right to freedom of expression and freedom of association are qualified rights. We must exercise our duties under the Act in light of users' and services' Article 10 and 11 rights and not interfere with these rights unless we are satisfied that to do so is prescribed by law, pursues a legitimate aim, is proportionate to the legitimate aim and corresponds to a pressing social need.

11.53 Removing a user's account from a service means removing that 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 in question for the duration of the removal. This effect also extends to other users, who will be unable to receive information shared by the relevant user via the removed account on the service in question. While a user whose account is removed in accordance with this measure is not necessarily prevented from creating a new account with the service from which they have been removed, their rights will still be affected by account removal due to the loss of their network and, in many cases, their content.

[1652] This is based on the high labour cost assumptions set out in Annex 5. It would be around £70 based on our low labour cost assumptions. We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS. We received some feedback on the general cost assumptions (e.g. salary assumptions) that are fed into these costs. We consider that feedback in Annex 5.

[1653] We recognise that setting up and/or devising an automated process for removing accounts found to be operated by or on behalf of a proscribed organisation would be more involved and require different ICT professionals' input.

11.54 It is unclear whether a proscribed organisation necessarily has rights to free expression or free association, given that (1) a terrorist organisation's purposes are fundamentally inconsistent with democracy and human rights, and (2) a terrorist organisation does not usually have legal personality.

11.55 In any event, if a proscribed organisation does have rights to free expression or free association, the effect of proscription on those rights has already been considered in the decision of the Home Secretary (as approved by Parliament) to proscribe that organisation. Therefore, the concerns set out in paragraph 11.53 regarding the effect of account removal on human rights do not arise for correctly identified proscribed organisations as they do for other users. Furthermore, it is clearly proportionate for the account of a proscribed organisation to be removed from a service for the duration of its proscription (which may be indefinite).

11.56 We recognise the importance of allowing service providers to consider the context of illegal content in deciding whether to remove an account, which will likely minimise risks of infringing users' right to freedom of expression. This concern was raised by CCAN, as outlined in paragraph 11.20.[1654] We agree that context must be a crucial consideration when a service provider is assessing whether to remove an account under this measure, and believe it permits providers to take into account such context by recommending that providers remove accounts where they have reasonable grounds to infer an account is run by or on behalf of a proscribed organisation.

11.57 We also recognise that this measure could lead to overreach or abuse targeted at peaceful dissenters, civil society and academic researchers.[1655] We have responded to these concerns in the ICJG: chapter 2 'Terrorism', explaining that the known identity of a user is relevant to determining whether content is illegal, including whether it amounts to a proscribed organisation or other terrorism-related offence. Where a provider decides content reviewed pursuant to this measure is not illegal for this reason, we would not expect it then to remove the account in question. As such, we would not expect the correct application of this measure to affect the free expression or association of peaceful dissenters, civil society and academic researchers.

11.58 In a similar vein, we also recognise that some proscribed organisations are elected governments, state-sponsored, or resourced to form quasi-governments, and the removal of these accounts could amount to interference with rights of users to receive information. This concern was raised by CCAN and is outlined in paragraph 11.17.[1656] In these cases, the accounts and associated online activity may provide public services that are essential for communities or constituents, such as news announcements or communication of information. Although we recognise that some proscribed organisations also perform governmental or administrative roles in other jurisdictions, we have explained above that any activity carried out on (or content produced on) accounts operated by proscribed organisations is very likely to amount to an offence in the UK. As explained in chapter 2 of this Volume: 'Content moderation' where the Act requires content to be taken down, this refers to taking it down for UK users.[1657] Accordingly, we maintain that the removal of

[1654] CCAN response to November 2023 Consultation, pp.2-3.

[1655] OxDEL response to November 2023 Consultation, p.8.

[1656] CCAN response to November 2023 Consultation, pp.2-3.

[1657] Section 8(3) of the Act states that the illegal content safety 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.

accounts operated by or on behalf of proscribed organisations is a proportionate action that will assist service providers in complying with their illegal content duties. The measure recommends that providers remove accounts operated by or on behalf of proscribed organisations, and not simply accounts that appear to be operated by users associated with those organisations (though providers will be under a duty to remove any illegal content posted by those accounts).

11.59 That said, we recognise that there is a risk to users' human rights if their accounts are incorrectly identified as being operated by or on behalf of a proscribed organisation and consequently removed. We also acknowledge importance of adequate and robust recourse mechanisms allowing users to appeal content or account removals. These concerns were reiterated by stakeholders during the November 2023 Consultation, including several recommendations to mitigate this issue, as outlined in paragraph 11.19.[1658] Ultimately, we believe the flexibility built into the measure recognises that the assessment carried out by a service provider will depend on the nature of its service and the specific circumstances of the relevant case. It allows service providers to consider the context and nature of an account (and any violative content spread via that account) to decide whether it has reasonable grounds to infer the account is operated by or on behalf of a proscribed group.

11.60 To further safeguard against incorrect identification and removal of accounts, the Act requires service providers to take appropriate action in response to certain complaints, including appeals by UK users against a decision to suspend them, ban them, or otherwise restrict their access to the service as a result of the user generating, sharing or uploading content that the provider considers to be illegal (see section 21(4)(d)). This obligation only applies if a service provider has made an illegal content judgement. We recognise that service providers' terms and conditions may consider a wider range of content to be 'terrorist content' than is defined in UK domestic law.

11.61 Given the broad nature of the offence of preparing a terrorist act, and the likelihood that content disseminated by a proscribed organisation account is illegal, we consider that any content takedown decision or account removal based on content related to a proscribed organisation is likely to be based on an illegal content judgement within the meaning of the Act, regardless of whether the provider considers the content to be illegal or to violate its terms and conditions. The complaints obligation would therefore apply, enabling UK users who believe their account or content has been wrongfully removed to appeal the decision.

11.62 Incorrectly removing a user's account would interfere with their rights to freedom of expression and association in the period between being removed and their appeal being considered. As explained above, such interference must be both prescribed by law and necessary for the achievement of a legitimate aim. To be considered necessary, the restriction must correspond to a pressing social need and be proportionate to the legitimate aim being pursued. Seeking to remove accounts operated by or on behalf of a proscribed organisation clearly serves the legitimate interests of national security, public safety, and the prevention of crime, and combatting terrorism is unarguably a pressing social need. We consider that the measure will be effective in reducing illegal harms perpetrated by proscribed accounts, including recruitment and support-gathering.

[1658] CCAN response to November 2023 Consultation, p.2; Centre for Competition Policy response to November 2023 Consultation, pp.2, 16; Federation of Small Businesses response to November 2023 Consultation, p.4; ISD response to November 2023 Consultation, p.12; Microsoft response to November 2023 Consultation, p.20.

11.63 Reversal of account removal following a successful user appeal will end the interference with that user's rights. Taking the above (from paragraph 11.52) into consideration – and noting that a service provider would have needed to establish the presence of a combination of factors before the account was removed – we consider any interference with user rights to be justified and proportionate in the circumstances.

11.64 Overall, while we acknowledge the potential interference of the measure with users' rights to freedom of expression and association where accounts are wrongfully removed, we consider the anticipated risks to be minimal and proportionate.

### Privacy

11.65 As explained in 'Introduction, our duties, and navigating the Statement', as well as chapter 14 of this Volume: 'Statutory tests', Article 8 of the ECHR sets out the right to respect for individuals' private and family life. An 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.

11.66 It is unclear whether proscribed organisations have a right to privacy because they do not usually have legal personality. However, we acknowledge that actions which may amount to an infringement of privacy may take place in relation to a user whose account is reviewed under this measure where the provider decides the account is not operated by or on behalf of a proscribed organisation. It must therefore be assumed that individual users have a right to privacy which may be infringed through the actions of the service provider.

11.67 We recognise that the implementation of this measure could have implications for individual users' rights to privacy (for example, when a service provider is reviewing content posted by a user). We recommend that service providers should consider whether an account may be run by or on behalf of a proscribed organisation when they have:

- identified content posted to the account that amounts to a proscribed organisation offence;
- identified content that is in breach of an equivalent standard as set out in their terms and conditions; or
- become aware of an account that may be operated by or on behalf of a proscribed organisation (including as a result of a report or complaint).

11.68 Assessing the account to determine if it is run by or on behalf of a proscribed organisation and any interference with privacy arising from that assessment is therefore only recommended where the provider has a reason to suspect that it may be a proscribed organisation account.

11.69 In many cases, a person will not have a reasonable expectation of privacy in publicly communicated content. In such cases, the right to privacy would not be engaged by the provider assessing the account. However, a person may have a reasonable expectation of privacy in publicly communicated content depending on the specific circumstances. Where this is the case, given the level of risk proscribed organisations pose to public safety and national security (as well as the fact that running such an account would likely be a criminal offence), we consider that the interference with the right to privacy which would be caused by the provider assessing a reasonable quantity of the content on the account to make a decision is likely to be proportionate.

11.70 As discussed in paragraph 11.35, we do not consider it practicable to specify precisely how much content a service provider should consider when making a decision on removal, as we expect that this will vary both between service providers and individual cases. However, we consider that recommending service providers take only a 'reasonably sized sample' of 'recent content' makes it clear that unless there is very little content produced via the account, it is not necessary for service providers to review all the content on the account. This will further ensure that any interference that affects user privacy is proportionate.

11.71 It is much more likely that a person would have a reasonable expectation of privacy in privately communicated content. As such, we have concluded that it would not a be proportionate interference with user privacy to expect service providers to review content communicated privately, unless they have explicit consent to review specific content.

11.72 Given these factors and the potential for this measure to mitigate the significant harm posed by accounts operated by or on behalf of proscribed organisations, we consider that any interference with a user's right to privacy to be proportionate.

### Data protection

11.73 We recognise that implementing this measure will mean that service providers have to process personal data when reviewing an account that they would not otherwise process and, to the extent they come to a view about the activity of individual users, may also create personal data. Providers will remain subject to applicable privacy and data protection laws when carrying out any such processing, including the principle of data minimisation, in determining how much content to review. Providers should refer to relevant guidance from the ICO.[1659]

## Who this measure applies to

11.74 This measure applies to all U2U services.

11.75 In reaching this decision, we have carefully considered whether to apply this measure to services that are small and low-risk for all kinds of illegal harm. The vast majority of services that are small and low-risk for all kinds of illegal harm are unlikely to host an account operated by or on behalf of a proscribed organisation. That said, in our November 2023 Consultation, we recognised the risk of proscribed organisations migrating to smaller services after their account has been removed from a larger service and noted that there is evidence of the use of smaller services for spreading terrorist content.[1660] [1661]

11.76 We believe that recommending all service providers remove proscribed organisations' accounts should help to mitigate this risk. If a small service were targeted by a proscribed organisation, there could be a delay in the provider finding its service is at medium or high risk of terrorism offences until it undertakes its next risk assessment. During that period, there would be a benefit from imposing this measure on the service, as opposed to only imposing it on services that are higher risk for terrorism offences.

11.77 Where services are not targeted with terrorist content, providers will only incur the costs of reading and understanding this measure. Given the limited costs and the severe nature of

[1659] Information Commissioners Office. Online safety and data protection. [accessed October 29, 2024].

[1660] In the first half of the 2010s, groups like the Global Islamic Media Front (GIMF), Al Qaeda, and ISIS had a significant presence on 'conventional' social media sites.

[1661] Amarasingam, A., Maher, S., and Winter, C.., 2021. How Telegram Disruption Impacts Jihadist Platform Migration. [accessed October 29, 2024].

the harm caused by proscribed organisations, we consider it proportionate to apply the measure to all services, including services that are small and low risk for all kinds of illegal harm.

## Conclusion

11.78 In view of the harm content generated, shared or uploaded by or on behalf of a proscribed organisation can cause, and the fact that such content is very likely to be illegal, we consider this measure to be proportionate. This is particularly the case given that the costs of the measure are only likely to be material in circumstances where services host significant numbers of proscribed organisation accounts.

11.79 When implementing this measure, service providers are likely to review content posted by a range of users, including users who are not involved with proscribed organisations, and there is some risk that they may incorrectly remove such accounts. This may therefore interfere with the right to freedom of expression, association, and privacy of the affected users. However, for the reasons set out in this chapter – and in particular the potential for this measure to mitigate the significant risk of harm posed by accounts run by proscribed organisations – we consider any interference would be proportionate.

11.80 Taking all of this into account, we have decided to go ahead with this measure as proposed in our November 2023 Consultation. This measure is included in our Illegal Content Codes of Practice for U2U services and is referred to as ICU H1. This measure is part of our Illegal Content Codes of Practice on terrorism.

# Option explored on banning user accounts that share CSAM

11.81 In the November 2023 Consultation, and as mentioned in paragraph 11.5, we considered proposing a measure that recommends U2U service providers block users where they are found to have uploaded or shared content relating to or facilitating the dissemination of child sexual abuse material (CSAM).

11.82 We decided not to consult on proposals at the time as we needed to work through the detail of the measure. We set out our intention to consult at a later date on a proposal on blocking users who are found to have uploaded or shared CSAM. To inform this work, we invited respondents to provide evidence to support our consideration of a future measure that would recommend banning users that spread CSAM.

## Summary of stakeholder feedback[1662]

11.83 A number of stakeholders provided evidence emphasising the seriousness of CSAM, including the harm it causes to users that are exposed to such content, and supported a measure to ban user accounts that spread CSAM.[1663] All responses received are detailed in

[1662] Note this list is not exhaustive, and further responses can be found in Annex 1.

[1663] Canadian Centre for Child Protection (C3P) response to November 2023 Illegal Harms Consultation, p.27; Centre for Competition Policy response to November 2023 Consultation, p.17; GeoComply Solutions response to November 2023 Illegal Harms Consultation, pp.11-15; Marie Collins Foundation response to November 2023 Illegal Harms Consultation, p.18; Match Group response to November 2023 Consultation, p.18; Scottish the Annex 1. They also reinforced our existing understanding that a range of service providers currently operate policies which involve banning users and accounts that spread CSAM.[1664] [1665]

## Our decision and reasoning

11.84 The submissions we received strongly reinforced the view that a CSAM banning measure would be proportionate. We will consult on a measure in this space in the further consultation we plan to publish in Spring 2025.

# Option explored on strikes and banning for illegal content

11.85 In the November 2023 Consultation, and as mentioned in paragraph 11.5, we also considered proposing a measure which would recommend that U2U service providers implement a strike and blocking system for users who post any type of illegal content or facilitate illegal behaviour.

11.86 We provisionally concluded that we could not make a recommendation for a single system for strikes and banning that would be suitable for all types of services and harms (though this does not preclude service providers from operating these systems as tailored to their services.) We also determined that more evidence would be required to ensure appropriate safeguards are in place to protect users' rights. For these reasons, we did not propose a measure which would recommend that service providers employ a strikes and blocking system to address accounts posting any type of illegal content.

Government response to November 2023 Illegal Harms Consultation, p.11; Welsh Government response to November 2023 Illegal Harms Consultation, p.5.

[1664] Name withheld 5 response to November 2023 Consultation, pp.14-15; Google response to November 2023 Consultation, pp.67-68; [✂]; Microsoft (confidential) response to November 2023 Consultation, p.20; Name withheld 4 (confidential) response to November 2023 Illegal Harms Consultation, p.10; [✂]; UK Interactive Entertainment (Ukie) response to November 2023 Illegal Harms Consultation, pp.29-30; WeProtect Global Alliance response to November 2023 Illegal Harms Consultation, pp.23-24.

[1665] We set out in the November 2023 Consultation our existing evidence base, which included: (1) TikTok stated that it issues permanent bans on first violation for "promoting or threatening violence, showing or facilitating child sexual abuse material (CSAM), or showing real-world violence or torture"; (2) Vimeo's Acceptable Use Community Guidelines state that "If we locate any content suspected of containing CSAM, we will immediately remove the account…Certain users may not use our services, regardless of their content. These are: gangs, hate groups, terror organizations, members of the foregoing"; (3) X stated that "in the majority of cases, the consequence for violating our child sexual exploitation policy is immediate and permanent suspension." It permanently suspends any accounts that violate its violent and hateful entities policy; (4) WhatsApp stated that it "ban[s] users when we become aware they are sharing content that exploits or endangers children."; (5) Name withheld 5 stated that it disables the user's account in the case of confirmed CSAM; and (6) Meta stated that it will "disable the user's account, Page or Community on Facebook, or the user's account on Instagram, after one occurrence" of child sexual exploitation content is detected.

## Summary of stakeholder feedback[1666]

11.87 Our analysis of consultation responses indicates that stakeholders are supportive of a measure that would apply to all illegal content or specified specific kinds of illegal harm.[1667] A number of respondents emphasised that recommending measures to remove accounts that post or share other types of illegal and harmful content would help foster a safer online environment. These submissions are detailed further in Annex 1.

## Our decision and reasoning

11.88 We have decided not to recommend a broad strike and banning measure in our Codes at this time.

11.89 Although responses to the November 2023 Consultation were supportive of such a measure, we did not receive new evidence to justify the inclusion of a measure that bans accounts spreading any illegal content or facilitating illegal behaviour.

11.90 As a result, our reasons for not recommending this measure remain the same as those communicated in our November 2023 Consultation: (1) there is no single system that would be suitable for all types of services and harms; and (2) more evidence is required to ensure that appropriate safeguards are in place to protect user rights.

# Assessment of verifying user identity

11.91 In the November 2023 Consultation, we considered the case for recommending that services deploy identity verification as a potential mitigation against the risk posed by users posting anonymously online. However, for the reasons set out in the November 2023 Consultation we decided not to consult on a measure.

11.92 We explained that our evidence of the efficacy of user verification in deterring illegal content was mixed, and we considered there to be important benefits to anonymity for some groups. There are also complex user rights implications associated with identity verification.

## Summary of stakeholder feedback[1668]

11.93 In response, a number of stakeholders requested identity verification measures to address other types of illegal harms.[1669] These are detailed further in Annex 1. Others expressed

[1666] Note this list is not exhaustive, and further responses can be found in Annex 1.

[1667] Cifas response to November 2023 Illegal Harms Consultation, p.18; [✂]; Local Government Association response to November 2023 Consultation, p.14; Monzo response to November 2023 Illegal Harms Consultation, pp.20-21; UK Finance response to November 2023 Illegal Harms Consultation, pp.2, 7, 11, 13.

[1668] Note this list is not exhaustive, and further responses can be found in Annex 1.

[1669] [✂]; Clean up the Internet (CUTI) – Proposal for a measure requiring platforms to offer their users options to verify their identity, response to November 2023 Illegal Harms Consultation, p.2; Community Security Trust and Antisemitism Policy Trust response to November 2023 Illegal Harms Consultation, p.14; Innovate Finance response to November 2023 Illegal Harms Consultation, pp.2, 11-12, 16; LoveSaid response to November 2023 Illegal Harms Consultation, p.15; [✂]; [✂]; OneID response to November 2023 Illegal Harms Consultation, pp.3-5; Philippine Survivor Network response to November 2023 Illegal Harms Consultation, p.10; UK Finance response to November 2023 Consultation, pp.2, 3, 13; Yoti response to November 2023 Illegal Harms Consultation, pp.17-18.

concerns about identity verification, focussing on the value of anonymity online and the privacy impact of identity verification.[1670]

## Our decision and reasoning

11.94 We have decided to confirm our proposal not to recommend an identity verification measure.

11.95 The November 2023 Consultation responses did not lead to any changes from our original position. We recognise that there is evidence that anonymity can give rise to risks of certain illegal harms. Set against this, anonymity also provides some significant benefits, particularly to marginalised groups. In light of these benefits, we are not persuaded that recommending compulsory identity verification in our Codes would be proportionate. The Act imposes a requirement for the providers of categorised services to give their users the option of verifying their identity and filter out content from non-verified accounts. We are currently developing our approach to the implementation of these duties.[1671] We note that, in its response, Clean up the Internet (CUTI) set out how an optional identity verification measure could work and proposed that we include such a measure in our illegal harms Codes.[1672] We intend to consider the case for this proposal as part of our work on the categorised services, so that we can take a holistic view on identity verification at that point. Following this, we will be able to consider the case for incorporating identity verification into other measures in the future.

[1670] Centre for Competition Policy response to November 2023 Consultation, p.17; Mid Size Platform Group response to November 2023 Illegal Harms Consultation, p.11; Name withheld 3 response to November 2023 Illegal Harms Consultation, p.18; Nexus response to November 2023 Consultation, p.18; Ofcom's Advisory Council for Northern Ireland response to November 2023 Illegal Harms Consultation, pp.9-10; Reddit response to November 2023 Illegal Harms Consultation, p.22.

[1671] More information about categorised services and Phase 3 can be found at Ofcom's approach to implementing the Online Safety Act.

[1672] CUTI – Proposal for a measure requiring platforms to offer their users options to verify their identity, response to November 2023 Consultation, p.2.

# 12. User Controls

## What is this chapter about?

This chapter sets out the user controls measures we are recommending, why we are recommending them, and to which user-to-user (U2U) services they should apply.

## What decisions have we made?

We are recommending the following measures:

| Number in our Codes | Recommended measure | Who should implement this |
|---|---|---|
| ICU J1 | Providers should offer every registered user options to **block and mute** other user accounts on the service. | Providers of large U2U services that:<br>• have user profiles and certain user interaction functionalities; and<br>• are at medium or high risk of one or more of the following kinds of illegal harm: grooming; encouraging or assisting suicide (or attempted suicide); hate; harassment, stalking, threats, and abuse; and coercive and controlling behaviour. |
| ICU J2 | Providers should offer every registered user the option of **disabling comments** on their content. | Providers of large U2U services that:<br>• enable users to comment on content; and<br>• are at medium or high risk of one or more of the following kinds of illegal harm: grooming; encouraging or assisting suicide (or attempted suicide); hate; harassment, stalking, threats, and abuse. |
| ICU J3 | Providers should provide information to help users understand why and how user profiles are labelled under notable user schemes and monetised schemes. They should also have and consistently apply clear internal policies on these schemes. | Providers of large U2U services that:<br>• are at medium or high risk of fraud or the foreign interference offence; and<br>• operate a notable user scheme or monetised scheme. |

## Why are we making these decisions?

The measures in this chapter, taken together, aim to ensure that service providers give users access to tools that allow them to determine the content they see on services, who can contact them and who can interact with them, and information that helps them to decide whether to engage with and trust content.

Enabling users to block other users can help them reduce the risk of encountering illegal content. In particular it can play an important role in helping users avoid harms such as harassment, stalking, threats and abuse, and coercive and controlling behaviour. Similarly, allowing users to disable comments can be an effective means of helping them avoid a range of illegal harms including harassment (such as instances of epilepsy trolling and cyberflashing) and hate. These offences are widespread and cause significant harm.

In light of the prevalence and impacts of the harms and the important role we consider the measures could play in tackling them, we consider that the benefits of these measures are sufficient to justify the costs we have identified. There is a degree of uncertainty about some of the costs. In order to ensure that we are acting proportionately, we have decided to target the measures at medium or high-risk large services.

Our measure relating to notable user and monetised profile labelling schemes (ICU J3) should increase user understanding of why profiles are labelled, enabling them to take this context into account when deciding whether to engage with content posted via the account in question. It will provide users with information to reduce the risk of them falling victim to foreign interference or fraud. We have made three minor amendments to the measure, which are set out in the relevant 'Our decision' section.

## Introduction

12.1 The Online Safety Act 2023 ('the Act') requires providers of regulated user-to-user ('U2U') services to take certain steps to reduce the risk of harm to users from illegal content. The requirements include taking proportionate measures relating to the design or operation of a service to mitigate and manage the risks of harm to individuals (section 10(2)). The Act states that one of the areas to which the duties apply is (where proportionate) "functionalities allowing users to control the content they encounter" (section 10(4)(f)).

12.2 The measures presented in this chapter aim to give users more control or understanding of the content they encounter and to give them tools to protect themselves from encountering illegal content.

12.3 In this chapter, we first address the feedback we received relating to our proposals for our measures on user blocking and muting (ICU J1) and disabling comments (ICU J2) together, as most respondents treated these measures together in their responses. We then address the feedback we received relating to our measure on notable user and monetised schemes (ICU J3).

## Measures on user blocking and muting, and disabling comments

12.4 In our November 2023 Illegal Harms Consultation ('November 2023 Consultation), we proposed that all providers of large U2U services that:

- identify as medium or high risk for any of the following harms: coercive and controlling behaviour; harassment, stalking, threats and abuse; hate; grooming; encouraging or assisting suicide;[1673]

[1673] The measure as consulted on in the November 2023 Consultation applied to services which are at medium or high risk of 'Encouraging or assisting suicide' and 'Encouraging or assisting serious self-harm'. We have now taken 'Encouraging or assisting self-harm' out of scope as we have made sure the harms groupings only include priority offences, consistent with Parliament's decision that they should be a priority.

- have user profiles; and
- have at least one of the following functionalities: user connections; posting content; or user communication (including but not limited to direct messaging and commenting on content),

should offer every registered user the option to block or mute other user accounts on the service (whether or not they are connected on the service), and the option to block all non-connected users (which we refer to as 'global blocking') (measure ICU J1).

12.5 We also proposed that providers of large U2U services that:

- identify as medium or high risk for any of the following harms: harassment, stalking, threats and abuse; hate; grooming; encouraging or assisting suicide;[1674] and
- enable users to comment on content,

should offer every registered user the option of disabling comments on their own posts (measure ICU J2).

12.6 We also asked for feedback on whether our measures on blocking and muting user accounts and disabling comments should include provisions how controls are made known to users.

12.7 We proposed these measures to reduce the risk of a user encountering illegal content by empowering them to block and mute other user accounts. These tools can play an important role in helping users to avoid harms such as harassment, stalking, threats and abuse, grooming, hate, encouraging or assisting suicide and coercive and controlling behaviour. Similarly, allowing users to disable comments can be an effective means of helping them to avoid a similar range of illegal harms.

## Summary of stakeholder feedback[1675]

12.8 There was broad support for both measures across a range of respondents.[1676] In addition to support for these measures, stakeholders highlighted several areas for further consideration:

[1674] The measure as consulted on in the November 2023 Consultation applied to services which are at medium or high risk of 'Encouraging or assisting suicide' and 'Encouraging or assisting serious self-harm'. We have now taken 'Encouraging or assisting self-harm' out of scope as we have made sure the harms groupings only include priority offences, consistent with Parliament's decision that they should be a priority.

[1675] Note this list in not exhaustive, and further responses can be found in Annex 1.

[1676] ACT The App Association (ACT) response to November 2023 Illegal Harms Consultation, p.17; Are, C. response to November 2023 Illegal Harms Consultation, p.16; Betting and Gaming Council response to November 2023 Illegal Harms Consultation, p.12; British and Irish Law, Education and Technology Association (BILETA) response to November 2023 Illegal Harms Consultation, p.16; Canadian Centre for Child Protection (C3P) response to November 2023 Illegal Harms Consultation, p.25; Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) response to November 2023 Illegal Harms Consultation, p.12; [✂]; Clean up the Internet response to November 2023 Illegal Harms Consultation, p.3; Community Security Trust and Antisemitism Policy Trust response to November 2023 Illegal Harms Consultation, p.14; Federation of Small Businesses response to May 2024 Consultation on Protecting Children from Harms Online, p.7; Global Network Initiative response to November 2023 Illegal Harms Consultation. p.17; INVIVIA response to November 2023 Illegal Harms Consultation, pp.23-24; LinkedIn response to November 2023 Illegal Harms Consultation, p.15; Local Government Association response to November 2023 Illegal Harms Consultation,

- feedback on how the measures work;
- prescriptiveness of the measures;
- placing the onus of safety on the individual;
- feedback on the effectiveness of the measures;
- costs of implementing the measures; and
- feedback on rights impacts.

12.9 We outline these in the following paragraphs.

### Feedback on how the measures work

12.10 Mid Size Platform Group said that blocking individuals in group messaging settings would be "impossible".[1677]

12.11 In its response to the May 2024 Protecting Children from Harms Online Consultation ('May 2024 Consultation'), the Canadian Centre for Child Protection ('C3P') said that the part of the blocking measure that allows users to block all unconnected users on a service should be applicable to children who have private accounts, and that it should prevent connection requests.[1678]

12.12 WhatsApp requested clarification on the scope of the equivalent disabling comments measure that we proposed as part of the May 2024 Consultation on Protecting Children from Harms Online.[1679]

12.13 We address these points in the 'How these measures work' section.

### Prescriptiveness of the measures

12.14 Several respondents said that the measures were too prescriptive:

p.13; Match Group response to November 2023 Illegal Harms Consultation, p.17; we note that Match Group made the same point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.6; Mencap response to November 2023 Illegal Harms Consultation, p.14; The National Society for the Prevention of Cruelty to Children (NSPCC) response to November 2023 Illegal Harms Consultation, pp.40-41; Nexus response to November 2023 Illegal Harms Consultation, p.17; Protection Group International response to November 2023 Illegal Harms Consultation, p.11; Refuge response to November 2023 Illegal Harms Consultation, p.20; Safe Space One response to November 2023 Illegal Harms Consultation, p.17; Segregated Payments Ltd response to November 2023 Illegal Harms Consultation, p.13; SPRITE+ (York St John University) response to November 2023 Illegal Harms Consultation, p.16; Stop Scams UK response to November 2023 Illegal Harms Consultation, p.16; The Cyber Helpline response to November 2023 Illegal Harms Consultation, p.18; UK Interactive Entertainment (Ukie) response to May 2024 Consultation on Protecting Children from Harms Online, p.50; Welsh Government response to November 2023 Illegal Harms Consultation, p.4.

[1677] Mid Size Platform Group response to November 2023 Illegal Harms Consultation, p.11. We note that Mid Size Platform Group made the same point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.12.

[1678] The Canadian Centre for Child Protection (C3P) response to May 2024 Consultation on Protecting Children from Harms Online, p.30.

[1679] WhatsApp response to May 2024 Consultation on Protecting Children from Harms Online, annex, p.13. This comment was made in relation to the equivalent proposed measure in our Protection of Children Codes but we consider this feedback to be also relevant to our Illegal Harms measure.

- Mid Size Platform Group said that we should be open to alternative user controls where there are practical challenges with implementing the proposed measures due to cost or service functionality.[1680]
- Google and techUK said that the measures are too prescriptive and may not be the best means for service providers to tackle harms.[1681] Google warned this could hinder providers from developing more effective compliance methods and suggested the Codes should be more flexible to allow for various current and future technological solutions to benefit from the safe harbour provisions.[1682]
- [✂] said that our measure on user blocking and muting should be expressed as an aim in the codes rather than explicitly requiring the introduction of specific user controls.[1683]
- Meta said that services should be able to develop proportionate solutions for blocking and muting user accounts, arguing that what constitutes 'proportionate' may vary by service.[1684]
- Snap and Pinterest, while broadly supporting the aims of the measures, said that their services largely met all or some of the measures through other means. They requested that greater flexibility be built into the measures.[1685] [1686]

12.15 We respond to these arguments in the 'How these measures work' section.

## Placing the onus of safety on the individual

12.16 Several respondents told us that while these types of user controls are helpful, the presence of these controls should not excuse services from taking a 'safety by design' approach by placing the responsibility of user safety primarily on users themselves.[1687]

12.17 We address this issue in the 'How these measures work' section.

[1680] Mid Size Platform Group response to November 2023 Consultation, p.11. We note that Mid Size Platform Group made the same point in response to the May 2024 Consultation, p.12.

[1681] Google response to November 2023 Illegal Harms Consultation, p.59; We note that Google made the same point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.40; techUK response to November 2023 Illegal Harms Consultation, p.27.

[1682] Google response to November 2023 Consultation, pp.2-3, 59.

[1683] [✂].

[1684] Meta response to November 2023 Illegal Harms Consultation, annex, p.15.

[1685] Snap said that the design of Snapchat prevents unconnected users from exchanging and viewing each other's content and gives users control around which comments on their content they choose to make public. Source: Snap response to November 2023 Illegal Harms Consultation, p.23.

[1686] Pinterest said its existing block function prevents blocked users from messaging, following, or commenting on the comment of blocking users, but does not stop blocked user content appearing in the blocking user's feeds. Source: Pinterest response to November 2023 Illegal Harms Consultation, p.9.

[1687] 5Rights Foundation response to November 2023 Illegal Harms Consultation, p.31; Institute for Strategic Dialogue response to November 2023 Illegal Harms Consultation, p.12; NSPCC response to November 2023 Consultation, p.41; Refuge response to November 2023 Consultation, p.20.

## Making the measures known to users

12.18 Several respondents agreed that the measures should include a recommendation that service providers make the controls known to users.[1688] [1689]

12.19 A number of stakeholders said that the information should be accessible:

- Mencap said that the information should be accessible to those with learning disabilities. [1690]
- INVIVIA made a general point in its response that user controls should be accessible to people of all ages and digital literacy levels.[1691]
- The National Crime Agency said that children should be given information when presented with options to block or mute other users. [1692]
- The National Society for the Prevention of Cruelty to Children (NSPCC) said that where this information is located on a service should be informed by where it would be most useful for users. [1693]
- Refuge said information about what the measures will do and how to use them should be provided in multiple languages and accessible formats.[1694]
- 5Rights Foundation said consideration should be given to the Information Commissioner's Office (ICO) Age Appropriate Design Code and the Institute of Electrical and Electronics Engineers (IEEE) Standard for an Age Appropriate Digital Services Framework.[1695]

12.20 Several other respondents cautioned that we should not be too prescriptive in how service providers should make these measures known to users, to allow flexibility for the different nature and functionalities of services.[1696]

12.21 We address all this feedback in 'How these measures work'.

12.22 Safe Space One disagreed with the inclusion of provisions on informing users but did not say why. [1697]

[1688] Are, C. response to November 2023 Consultation, p.16; BILETA response to November 2023 Consultation, p.16; C3P response to November 2023 Consultation, p.26; [✂]; Clean up the Internet response to November 2023 Consultation, 2023, p.5; INVIVIA response to November 2023 Consultation, p.24; Local Government Association response to November 2023 Consultation, p.13; Nexus response to November 2023 Consultation, p.18; OnlyFans response to November 2023 Illegal Harms Consultation, p.9; Protection Group International response to November 2023 Consultation, p.11; Segregated Payments Ltd response to November 2023 Consultation, p.13; The Cyber Helpline response to November 2023 Consultation, p.18.
[1689] Electronic Frontier Foundation support the intention of the proposal but do not think it should be law. Source: Electronic Frontier Foundation response to November 2023 Illegal Harms Consultation, p.18.
[1690] Mencap response to November 2023 Consultation, p.14.
[1691] INVIVIA response to November 2023 Consultation, p.24.
[1692] NCA response to May 2024 Consultation on Protecting Children from Harms Online, p.15.
[1693] NSPCC response to November 2023 Consultation, p.45.
[1694] Refuge response to November 2023 Consultation, p.20.
[1695] 5Rights Foundation response to November 2023 Consultation, p.31.
[1696] ACT The App Association response to November 2023 Consultation, p.17; LinkedIn response to November 2023 Consultation, p.15; Match Group response to November 2023 Consultation, p.17; Meta response to November 2023 Consultation, annex, p.15; Snap response to November 2023 Consultation, p.23.
[1697] Safe Space One response to November 2023 Consultation, p.17.

## Feedback on the effectiveness of the measures

### Feedback on effectiveness of the blocking and muting measure

12.23 [✂] questioned the benefit of allowing a user to hide all content from another user, so that they could not view any such content even if they wished to, for example if they chose to search for it. It queried how this type of functionality would address the relevant harms.[1698]

12.24 Some stakeholders also said it was not clear that the measure would have benefits or be effective if applied to certain services:

- Booking.com said that requiring blocking and muting functionalities would be disproportionate in circumstances where the interactions between users on its service were limited.[1699]
- Google expressed concern that the measures applied equally to social media services and services that have minimal social functionality.[1700] In its response to our May 2024 Consultation, Google said that the blocking and muting measure should not be applicable to all services that allow the posting of content.[1701]
- [✂] said that the ability to block users does not translate to video-sharing platforms due to users' limited ability to interact with one another.[1702]
- Regarding muting, Pinterest said this function would have minimal benefit on its service as its functionalities substantially reduces the likelihood of users viewing content of other users that are not being followed.[1703]

12.25 We address this feedback in the 'Benefits and Effectiveness' section. We also mention Booking.com and Google's feedback in the 'Who these measures apply to' section.

### Feedback on the effectiveness of the disabling comment measure

12.26 Snap, while being broadly supportive of the recommendation for services to allow users to disable comments on their content, said that its current functionality (which requires users to manually approve all inbound comments before they can appear publicly) achieved stronger protection than our disabling comments measure.[1704]

12.27 We address this in the 'Benefits and effectiveness' section.

## Costs and implications of implementing the measures

12.28 While no respondents quantified the costs of implementing the measures, we received feedback from stakeholders related to the potential high level of costs that these measures could entail for service providers. Some stakeholders said that all or parts of the measures were inappropriate for some services because they were not compatible with the purpose or proper functioning of the service, or would have limited or no benefits.

[1698] [✂].

[1699] Booking.com response to November 2023 Illegal Harms Consultation, p.21.

[1700] Google response to November 2023 Consultation, p.59.

[1701] Google response to May 2024 Consultation, p.40.

[1702] [✂].

[1703] Pinterest response to November 2023 Consultation, p.9.

[1704] Snap response to November 2023 Consultation, p.23. We note that Snap made the same point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.26.

- Mid Size Platform Group highlighted that our measures may not be straightforward to implement and could be resource-intensive. [✂], leading to technical issues in implementing them and risking disruption to user experience.[1705]
- [✂].[1706]
- Similarly, Global Network Initiative (‘GNI’) encouraged us to carefully consider the costs these measures could impose on providers, “especially when they are required on the part smaller or non-commercial service”.[1707]
- Match Group said that a blanket rule requiring providers to allow users to block all unconnected users on a service does not recognise the fundamental differences in the purpose of services, and this functionality does not translate to dating services.[1708]
- Wikimedia Foundation further said that the measures would not be appropriate for Wikipedia Talk pages given their reliance on editors being able to have debates to agree on content for Wikipedia articles.[1709]
- Reddit said that giving users of a discussion forum the ability to turn off comments on their content is “nonsensical” and would threaten the integrity of such services.[1710]
- An individual respondent did not support the measures, saying they were not “desirable or appropriate for all services”.[1711]

12.29 We consider these concerns in the ‘Costs and risks’ section. We also respond to GNI, Match Group, Wikimedia Foundation and Reddit’s feedback under the ‘Who these measures apply to’ section. We also address Wikimedia Foundation and Reddit’s from a rights angle under the section titled ‘Rights impact’.

### Feedback on freedom of expression considerations

12.30 Google said that blocking functionality raises freedom of expression considerations.[1712]

12.31 We address this in the section ‘Rights impact’.

### Feedback on who these measures apply to

12.32 Several respondents said the measures should be extended to all sizes of service.

- BT Group similarly argued that the measures appear to recommend basic functionalities that enable users to have control over their online experiences and reduce the risk of encountering harm, and therefore should be extended to providers of smaller services.[1713]
- Snap argued that failure to recommend these measures for all services would allow irresponsible design to become embedded at an early stage in a service’s product design

[1705] Mid Size Platform Group response to November 2023 Consultation, p.11.
[1706] [✂].
[1707] Global Network Initiative response to November 2023 Consultation, p.17.
[1708] Match Group response to November 2023 Consultation, p.17. We note that Match Group made the same point in response to the May 2024 Consultation, p.6.
[1709] Wikimedia Foundation response to November 2023 Illegal Harms Consultation, pp.33-34.
[1710] Reddit response to November 2023 Illegal Harms Consultation, pp.9-10 and 23.
[1711] Dwyer, D., response to November 2023 Illegal Harms Consultation, p.9.
[1712] Google response to November 2023 Consultation, p.65.
[1713] BT Group response to November 2023 Illegal Harms Consultation, p.2.

lifecycle.[1714] Snap also argued that recommending this measure exclusively to providers of large services risked giving a competitive advantage to providers of smaller services who would be exempt from implementing the measure, and therefore incur lower costs.[1715]

- Age Verification Providers Association and VerifyMy argued that the measure should be in place for child users wherever there are risks to children, including on smaller services.[1716]
- Some respondents said that the measures should be expanded to providers of smaller services. The Board of Deputies of British Jews, and UK Safer Internet Centre argued that these measures should not be limited to providers of large services, but rather should apply to all services regardless of size.[1717]

12.33 We address these comments in the section 'Who these measures apply to'.

## Our decision

12.34 We have decided to confirm the measures largely as we proposed in the November 2023 Consultation with two changes:

- We have added a provision that recommends providers make these measures known to users.
- We have also clarified the scope of the global blocking part of the measure by stating that it only applies to services that have user connection functionality.

12.35 The full text of the measures can be found in our Illegal Content Codes of Practice for U2U services on CSEA and other duties, in which they are referred to as ICUJ1 and ICUJ2.

## Our reasoning

### How these measures work

#### Measure on user account blocking and muting – blocking

12.36 In our November 2023 Consultation, we proposed that providers of large services that identify certain risks and have specific functionalities should offer every registered user the option to block or mute other user accounts on the service (whether or not they are connected on the service), and the option of global blocking.

12.37 This measure is designed to offer users the option to block individual connected or unconnected user accounts or all unconnected user accounts, and to mute other individual user accounts on the service.[1718]

[1714] Snap response to November 2023 Consultation, p.23.

[1715] Snap response to November 2023 Consultation, p.23.

[1716] Age Verification Providers Association response to November 2023 Illegal Harms Consultation, p.3; VerifyMy response to November 2023 Illegal Harms Consultation, p.13.

[1717] Board of Deputies of British Jews response to November 2023 Illegal Harms Consultation, p.3; Snap response to November 2023 Consultation, p.23; UK Safer Internet Centre response to November 2023 Consultation, p.12. We note that UK Safer Internet Centre made the same point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.37.

[1718] In the November 2023 Consultation, we included a definition of "user accounts" in the Codes of Practice. We have removed this definition on the basis that it is a commonly used and widely understood term which does not need to be defined.

12.38 ‘Block’ is commonly used terminology which refers to a user tool provided by U2U services that enables users to block connected or non-connected individual accounts or all non-connected accounts across the service.[1719] Here, in the context of User A blocking User B’s account, it means that:[1720]

- User B cannot send direct messages to User A and vice versa.
- User A will not encounter any content posted by User B on the service (regardless of where on the service it is posted) and vice versa, including but not limited to (1) reactions to and ratings of content and (2) content originally posted by user B which is subsequently posted by another user.[1721]
- If User A and User B were connected, they will no longer be connected.

12.39 In the context of User A blocking all non-connected user accounts on the service (global blocking), it means that:

- Non-connected users cannot send direct messages to User A (and vice versa).
- User A will not encounter the content of non-connected users and vice versa, including but not limited to (1) reactions to and ratings of content and (2) content originally posted by non-connected users which is subsequently posted by another user.

12.40 In response to C3P’s feedback, we are recommending the blocking measure (including individual blocking and blocking all unconnected users) for all users on a service, regardless of the privacy settings a user has set up on their user profile. This measure should also result in blocked users being unable to send the blocking user a connection request as they should not be able to find any of the blocking user’s profile information on the service, given this constitutes content posted by the blocking user under the definition in the Codes.

12.41 In response to stakeholder feedback, we agree that blocking individual users in group chat settings could cause user experience issues for all users in the group chat. Our measure recommends that the blocking user does not encounter any content generated, uploaded or shared by a blocked user on open channels of communication.[1722] It therefore does not apply to content that a blocking user may encounter through closed channels of communication such as group chats (with the exception of direct messages, as expressly stated in the measure). We note in our Risk Assessment Guidance that ‘group messaging’ takes place in a closed setting.

[1719] We have clarified the scope of the global blocking aspect of the measure to make clear that it applies only to services that have user connection functionality. Please see ‘Who these measures apply’ to for further information.

[1720] The effects set out in this, and the next paragraph describe the effect of blocking on the account through which User A has blocked User B’s account, and the account through which User B has been blocked by User A. For simplicity, we refer only to “users” rather than “user accounts” when discussing the action and effects of blocking in this chapter.

[1721] In this context, “Content originally posted by User B” relates to the entire content of their post. If User B were to post a link to a news article, the content includes the link to the news article and the information that shows User B used their user profile to make the post. If another user (User C) were to repost this content in a way that included the link and the information showing User B used their user profile to post this first, the blocking functionality should prevent User A from seeing User C’s post. If User C posted the link to the news article alone – without the accompanying information to show that User B originally posted it from their user profile – the blocking functionality should not prevent User A from seeing User C’s post. This also applies to global blocking.

[1722] See the definition of “content posted” in the U2U Illegal Content Codes.

### Measure on user account blocking and muting – muting

12.42 ‘Mute’ is commonly used terminology to refer to a user tool widely provided by U2U services that enables users to mute individual connected or non-connected user accounts on the service. Here, in the context of User A muting User B’s account, it means that User A will not encounter any content posted by User B on the service, including:[1723]

- reactions to and ratings of content by User B; and
- content originally posted by User B which is subsequently posted by another user (unless User A visits the user profile of User B, in which case User A will experience User B’s profile as if they had not muted them).

12.43 Muting is a ‘softer option’ than blocking as it allows the muted user to see the muting user’s content and leaves channels of communication open (for example, if User B were muted then they would still be able to direct message User A; if blocked, they would not be able to do so).

### Measure on disabling comments

12.44 In our November 2023 Consultation we proposed that providers of large U2U services that identify certain risks, and enable users to comment on content, should offer every registered user the option of disabling comments on their own posts. This option should be available when users first post content (so that they can prevent any comments at all), and after they have posted content (so that they can change their minds and turn comments off).

12.45 ‘Commenting on content’ is a functionality that allows users to reply to content or to post content in response to another piece of content that is visually accessible directly from the original content without navigating away from it.

12.46 WhatsApp requested clarification on the scope of the equivalent disabling comments measure that we proposed as part of the May 2024 Consultation on Protecting Children from Harms Online.[1724] This measure is designed to offer users the option of preventing other users of a service from commenting on content they have posted on open channels of communication.[1725] It therefore does not apply to closed channels such as private messaging and group chat settings.

### Prescriptiveness of both measures

12.47 As set out in the ‘Summary of stakeholder feedback’ section (see paragraphs 12.14 and 12.20), we received some feedback about the measures being too prescriptive, and that we should be open to alternative user controls to protect users from the harms that these measures seek to address.

[1723] The effects set out in this, and the next paragraph describe the effect of muting on the account through which User A has muted User B’s account, and the account through which User B has been muted by User A. For simplicity, we refer only to “users” rather than “user accounts” when discussing the action and effects of muting in this chapter.

[1724] WhatsApp response to May 2024 Consultation, annex, p.13. This comment was made in relation to the equivalent proposed measure in our Protection of Children Codes but we consider this feedback to be also relevant to our Illegal Harms measure.

[1725] As set out in the Codes, “posting content” relates to users generating, uploading or sharing content on open channels of communication. As the measure includes this definition, closed channels of communication are out of its scope.

12.48 As outlined in paragraph 12.14, some stakeholders suggested that they already achieve the outcomes of the measures through other means, while others argued the measures are too prescriptive, supporting flexibility to account for different service functionalities.

12.49 We acknowledge that many providers already offer similar functionalities on their services. We explain in the 'Our approach to developing Codes measures' chapter that we are required to have regard to several principles when setting our Codes.[1726] These are that the measures contained must be sufficiently clear, and at a sufficiently detailed level, that providers understand what those measures entail in practice, and that the measures for each kind and size of service must be proportionate to our assessment of the risk of harm they present.

12.50 We also set out in the chapter our awareness that there are many ways to 'design' a measure, some of which focus more on outcomes, some which are more prescriptive, and some high level. Our decisions for each measure flow from the principles above, our evidence base, and the results of our impact assessments.

12.51 We have carefully considered all stakeholder feedback and, based on our current evidence, we conclude at this time that these measures are the most effective and proportionate way to tackle the harms identified. Consistent with the principles that the Codes must follow, we consider them to strike the right balance as regards clarity and detail. As explained in the rest of this chapter, we consider them to be proportionate and technically feasible for the different types and sizes of services to which they apply, and proportionate to the risk of harm presented by those services. We set out the evidence relating to the effectiveness of these measures in the 'Benefits and effectiveness' section below.

12.52 We note that the Act provides flexibility for those who do not wish to adopt Codes measures, allowing them to adopt alternative measures to meet their duties. As set out in Volume 3: chapter 3: 'Ofcom's enforcement powers' and our 'Record-Keeping and Review Guidance', providers who wish to take alternative measures may do so as long as they record how this complies with the safety duties and their duties in relation to freedom of expression and privacy.

#### Placing the onus of safety on the individual

12.53 As explained in the 'Summary of stakeholder feedback' section, multiple respondents said that, while user controls are helpful, the onus should be on service providers to protect their users through 'safety by design' features.[1727] We agree with this feedback, which is why we have a diverse suite of measures. Some of these measures place an expectation for proactive action on service providers, while others provide users with options for protecting themselves from harmful content and making safer choices.[1728] The two measures described in this section of this chapter are not designed to move the onus of safety from providers to individuals, but rather are a complementary part of the wider package of measures.

[1726] Schedule 4 of the Act.

[1727] 5Rights Foundation response to November 2023 Consultation, p.31; Institute for Strategic Dialogue response to November 2023 Consultation, p.12; NSPCC response to November 2023 Consultation, p.41; Refuge response to November 2023 Consultation, p.20.

[1728] This is in line with the Act, which expressly refers to 'functionalities allowing users to control the content they encounter' as an area in which providers must take measures, if proportionate, to comply with their illegal content safety duties.

12.54 User controls are designed to provide an additional safety net for users who encounter illegal harms, as it is not realistic for all service providers to proactively identify all illegal material before any user might encounter it. Content moderation processes may not be fully effective against the harms this measure is designed to combat, including because of the nuance of how context-dependent content may constitute illegal harm. For example, content that can cause harm can include highly personal information, such as a stalker posting an image of a victim's front door to intimidate them.[1729] While user reporting and complaints give users the ability to alert service providers to illegal content, supplementary blocking and muting and disabling comments functions will help them to protect themselves from further harm immediately while the service investigates their complaint.

12.55 We discuss the benefits of allowing users to avoid encountering blocked users' content in the 'Benefits and effectiveness' section.

### Making the measures known to users

12.56 In the November 2023 Consultation, we asked stakeholders' views about whether the proposed measures on blocking and muting and disabling comments should include additional requirements for how these controls are made known to users.[1730]

12.57 Having considered feedback from stakeholders (see paragraphs 12.18-12.20), we have added a provision which recommends that providers give users information regarding these measures. This information must be easy to find and comprehensible based on the likely reading age of the youngest individual permitted to use the service without the consent of a parent or guardian.

12.58 Stakeholders also raised the importance of the information being accessible to those with learning disabilities or disabled people. We are not prescriptive about how this should be provided and therefore at this stage do not think it appropriate to specify particular ways in which it should be accessible to disabled people. We would expect approaches to accessibility to vary from service to service, subject to their services features and design and on that basis recommend providers are best placed to decide how to ensure information is accessible to disabled people. However, providers should consider their obligations under other relevant legislation (for example, the Equality Act 2010) and, where relevant, appropriate guidance.[1731] One stakeholder said information about what the measures will do and how to use them should be provided in multiple languages and accessible formats. We have not been prescriptive about the language in which service providers should give information to users regarding these features. We expect service providers would determine which language(s) would best suit their userbase.

12.59 We set out the benefits of this approach in the 'Benefits and effectiveness' section and outline the costs to service providers in the 'Costs and risks' section.

### Interaction with Children's Safety Codes

12.60 Our proposed Children's Safety Codes of Practice for U2U services recommends equivalent blocking and muting and disabling comments measures for certain services. It also includes a proposed recommendation (numbered PCU E1) that age-appropriate support materials be provided for children (including explanations for the adults who care for them) to ensure

[1729] Refuge, 2021. Unsocial Spaces. [accessed 22 October 2024].

[1730] "Consultation: Protecting people from illegal harms online", Ofcom, 2023, Volume 4, Chapter 20, p.281.

[1731] For example, World Wide Web Consortium's (W3C) Web Content Accessibility Guidelines (WCAG).

that they can understand the tools and know how to use them to mitigate the risk and impact of encountering harmful content.[1732] Subject to our final decisions relating to the Children's Safety Codes, services in scope of measure PCU E1 and our Illegal Content Codes measures ICU J1 and ICU J2 will likely be deemed to have met the recommendation in our Illegal Content Codes measures to give users information about the tools, by complying with the PCU E1 measure. This will reduce the burden on services that fall in scope of both sets of measures.

## Benefits and Effectiveness

12.61 Offences such as coercive or controlling behaviour, harassment, stalking, threats and abuse, hate, grooming and encouraging or assisting suicide take place online and cause significant harm. In this section, we discuss how damaging these harms are to users, how they manifest online, and why we think our measures will help tackle these harms on services of all types.

12.62 These offences can have significant effects on their victims and survivors. Each offence can have negative psychological impacts, including loss of confidence, anger and aggression, increased feelings of self-blame and shame, and lack of personal trust and isolation. Additionally, some offences, such as hate and coercive and controlling behaviour, can lead to financial harm, while others, such as encouraging or assisting suicide, can lead to physical harm and death. More details about the effects of each harm on their victims and survivors can be found in the Register of Risks ('Register').[1733]

12.63 We received feedback from several service providers that all or part of these measures do not translate to the functionalities on their services.[1734] However, we are aware of the relevant harms taking place on a variety of service types and we know that a wide range of functionalities, including direct messaging, livestreams, posting content and commenting on content can be risk factors.[1735] We also know that user controls such as blocking and muting functionalities are widely used by users to protect themselves from harm.[1736] Having carefully considered the responses, we maintain that this is the most effective and proportionate way to tackle these harms.

### Measure on user blocking and muting

12.64 For services which are designed primarily to promote user interaction through direct contact, the relevant harms often take place through such contact, for instance via direct messaging or commenting on content. Blocking functionality can help users protect themselves. For instance, in cases of cyberstalking, blocking communications from the offending user account can be one of the most effective methods of protection.[1737]

[1732] For further information, see "Protection of Children Code of Practice for user-to-user services", Ofcom, 2024, proposal PCU E1, pp.37-38.

[1733] For more details, see Register of Risks chapters titled 'Controlling or coercive behaviour', 'Harassment, stalking, threats and abuse', 'Hate', 'Child Sexual Exploitation and Abuse' (specifically the section on grooming) and 'Encouraging or assisting suicide'.

[1734] Booking.com response to November 2023 Consultation, p.21; Google response to November 2023 Consultation, p.59; Pinterest response to November 2023 Consultation, p.9.

[1735] See the relevant risk factors for each kind of illegal harm in our Risk Assessment Guidance and Risk Profiles.

[1736] 66% of respondents to a 2021 study from Thorn reacted to a harmful online experience by blocking the user and 27% muted the user. Source: Thorn, 2021. Responding to Online Threats: Perspectives on Disclosing, Reporting, and Blocking. [accessed 22 October 2024].

[1737] Tokunaga, R. S. and Aune, K. S., 2017. Cyber-Defense: A Taxonomy of Tactics for Managing Cyberstalking. Journal of Interpersonal Violence, 32 (10). [accessed 24 October 2024].

Similarly, in cases of abuse, limiting contact with an abusive account by blocking it can help users protect themselves from harmful behaviour.[1738]

12.65 Regarding [✂] comment (paragraph 12.23) about the rationale for preventing users from seeing each other's content, we consider this to be an important safeguard against the relevant offences.[1739] In cases of stalking, a user's content can be monitored to learn more about their preferences, activities, and whereabouts, and while monitoring content in itself is not an offence, it can facilitate stalking behaviour.[1740] Grooming offences often begin with the perpetrator identifying a child on a service by viewing the information in their user profile, such as their name, age, location and profile picture.[1741] Similarly, coercive or controlling behaviour can take the form of partner surveillance, with evidence showing that perpetrators of this offence utilise second and third degree connections to gain visibility of a target's user profile without connecting with them directly.[1742] In each of these instances, victims or survivors would be provided protection by the perpetrator being unable to find or view their content anywhere on the service.

12.66 We nevertheless recognise [✂] argument that there may be circumstances where a user blocks another user (for instance to avoid direct interactions) but still wishes to view certain content that the blocked user has posted on the service.

12.67 As these measures are about user choice, they are intended to empower users to consider their safety and how they can effectively protect themselves while making decisions about their experience on the service. In that context, there are options available to users to take action that is consistent with their risk appetite:

- The blocking user may choose to unblock the blocked user. These measures are designed to give users control, and we expect users should be able to choose when to switch blocking on and off.
- Users may wish to make use of the mute function, rather than the block function, which still allows them to see the content of the user they have muted when they actively choose to look for it.

12.68 There are also options available to service providers, as long as they offer users the blocking and muting options described in our measures:

- Service providers may choose to offer extra blocking options, in addition to the forms of blocking and muting recommended in our measures. For instance, providers of services where content discovery is an important part of the user experience may wish to offer users an additional option to block a user on a certain functionality (e.g. direct messaging), while still allowing them to encounter the blocked user's content elsewhere

[1738] Pen America, Online harassment Field Manual; *Blocking, Muting and Restricting*. [accessed 24 October 2024].

[1739] [✂].

[1740] Evidence shows that victims and survivors of stalking are likely to have had their activities monitored on social media services; the US Department of Justice in 2019 found this to be true of 31.9% of stalking victims and survivors. Source: US Department for Justice (Morgan, R. and Truman, J.), 2022. Stalking Victimization, 2019. [accessed 22 October 2024].

[1741] Quayle, E., Allegro, S., Hutton, L., Sheath, M. and Lööf, L., 2014. Rapid skill acquisition and online sexual grooming of children, *Computers in Human Behavior*, 39, pp.368-375. [accessed 6 November 2024].

[1742] Tseng, E., Bellini, R., McDonald, N., Danos, M., Greenstadt, R., McCoy, D., Dell, N., Ristenpart, T., 2020 The Tools and Tactics Used in Intimate Partner Surveillance: An Analysis of Online Infedelity Forums. [accessed 6 November 2024].

on the service. This would provide users with more options to decide how they experience the service, depending on the kinds of illegal harms they are encountering.[1743]

- Service providers may wish to inform users if their blocking choices can impact how they experience specific aspects of the service. As noted above, these measures are about giving users control over their experience, and we expect users to be able to turn blocking on and off. Providers may therefore also choose to give users the option to turn off blocking or muting at certain points of the user journey, or to switch to alternative forms of blocking if the provider has decided to offer these. For instance, we are aware that some online gaming services use randomised matching of users for certain games. If a blocking user and blocked user were randomly selected to join the same game, the provider could choose to offer the blocking user the option not to enter the gameplay, or to temporarily unblock the blocked user and enter the gameplay. A provider may decide that this temporary change to blocking should apply only to the extent necessary to enable gameplay, with protections still in place where feasible, such as blocks on voice chat and written messages.

12.69 In response to stakeholders who questioned the effectiveness of this measure for services where users' interactions are limited (see paragraph 12.24), we note that the harms targeted by this measure can take place through means that do not involve direct contact with the victim or survivor. For instance, evidence shows that harassment and stalking can take the form of public humiliation, where content about an individual is posted through open channels of communication.[1744] Similarly, a UK-based qualitative study found that content encountered outside of direct user interactions is a risk factor in encouraging suicide. Among the patients involved (which the study characterises as a higher severity group of self-harm patients with a history of suicidal behaviour), most had avoided generating online dialogue and instead preferred to observe others' posts on methods of harm to gain insight into experiences or decide on details of implementation.[1745] In situations where patients are in recovery, allowing the blocking user to block all content from the blocked user, in addition to direct contact from the blocked user, can help users protect themselves from harm.

12.70 Similarly, our understanding of the way harms captured by this measure manifest means that, in response to Pinterest's concern on the effectiveness of muting,[1746] we maintain that it is important to offer users a choice of blocking or muting other user accounts. Blocked users may discover they have been blocked (if, for example, they try to engage with the blocking user's content and cannot find it) whereas muted users cannot discover that they

[1743] This is not specifically recommended in our Codes, as the relevant options would depend on the characteristics of a service, while additional blocking options may also not be effective or proportionate in all cases.

[1744] Social media posts have been identified as the most common trigger for harassment. Source: Gosse, C., Veletsianos, G., Hodson, J., Houlden, S., Dousay, T.A., Lowenthal, P.R., and Hall, N., 2020. The hidden costs of connectivity. [accessed 22 October 2024].

[1745] This study refers to suicidal behaviour and self-harm. Whilst we recognise that encouraging self-harm is not targeted by the measure, this evidence corroborates evidence around other relevant harms, such as encouraging suicide which can sometimes follow similar patterns of behaviour. 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 22 October 2024].

[1746] Pinterest response to November 2023 Consultation, p.9.

have been muted in the same way. This can be advantageous in instances where a user is concerned that illegal behaviour could be escalated if another user were to discover that they have chosen to no longer see their content.[1747] Similarly, it can be useful in cases of harassment, where blocked users can sometimes create new fake accounts to circumvent a block they have discovered.[1748]

12.71 Broadly, regarding feedback outlined that this measure would be less effective for services with limited social interactions, we note such services would not necessarily have medium or high risk of one of the relevant harms.[1749] However, if a service does have medium or high risk of one of the relevant harms, we maintain that the provider should implement these measures to provide protection to users.

12.72 In view of these points, we consider that our blocking and muting measure will play an important role in combatting the relevant harms and will reduce victims' and survivors' exposure to such harms online. Given the prevalence of these harms and the severity of the impact they have, we consider that this will deliver significant benefits.

### Blocking all unconnected accounts on services without user connections

12.73 The global blocking part of the measure is intended for services that have user connection functionality. We recognise that providers of services without a user connection functionality may have interpreted the measure as implying that all user accounts on the service are 'unconnected' to the blocking user, meaning that a user using global blocking would experience the service as if they were the only user on it. This was not the intention of the measure. An option to block all other users may make such a U2U service unappealing or unusable to some users, and may therefore see little uptake. For these services we therefore recommend offering users the ability to block or mute other individual users, but do not consider it proportionate to recommend offering a global blocking feature, given its limited likely effectiveness.

12.74 This part of the measure can enhance protection from harm and empower users to make safer choices on these services in a range of circumstances. For instance, child users can be targeted by non-connected accounts to initiate conversation with the intent to groom.[1750] Allowing users to block all unconnected accounts can also be particularly useful in circumstances where perpetrators persistently try to circumvent individual blocks by creating fake accounts to contact their victim.[1751]

[1747] Refuge found that 15% of the women survivors responding to its survey said abuse worsened when they reported the perpetrator or took an action to mitigate the abuse, such as blocking the perpetrator online. Refuge, 2021. Unsocial Spaces. [accessed 22 October 2024].

[1748] Individuals who stalk, harass, or threaten others online are known to sometimes run multiple accounts when interacting with their victims. If one account is reported and banned, they can seamlessly move to another. Source: UK Home Office, 2021. Anonymous or multiple account creation: improve the safety of your online platform. [accessed 22 October 2024].

[1749] Booking.com response to November 2023 Consultation, p.21; Google response to November 2023 Consultation, p.59; Pinterest response to November 2023 Consultation, p.9.

[1750] A study by Kloess *et al.* found an example of a perpetrator randomly adding children to initiate contact with them. Source: Kloess, J. A., Hamilton-Giachritsis, C. E. and Beech, A. R., 2019. Offence Processes of online sexual grooming and abuse of children via internet communication platforms, *Sexual Abuse*, 31(1), pp.73-96. [accessed 29 October 2024].

[1751] The practice of creating fake accounts to cause harm has been observed in cases of technology-enabled domestic abuse, stalking, coercive control, and encouraging or assisting suicide. Sources: Refuge, 2021.

12.75 Regarding grooming, we expect that our recommended measure about safety defaults for child users (ICU F1), that restricts direct messages between non-connected users and child users will still protect child users from the risks of grooming on relevant services without user connections.[1752]

### Measure on disabling comments

12.76 Comments sections on U2U services can be a significant source of harm.[1753] [1754] While we do not have access to data or evidence about how widely comment disabling tools are used by users of U2U services, or how effective they consider this to be as a means of reducing the risk of exposure to illegal content, our view is that this measure will provide effective protection against harm for all users who choose to use it. It does this by giving users the choice to either:

- take immediate and comprehensive action to protect themselves from harm by preventing all other users from commenting on their content at the point of uploading the content to the service; or
- reduce their exposure to illegal content by preventing further comments on their content, by turning off comments at any point after the content has been uploaded to the service.

12.77 Giving this control to users is important because several of the relevant harms for this measure are nuanced and highly personal, and therefore users will benefit from being able to use this tool to protect themselves from illegal harms.

12.78 While the evidence suggests that harm committed through comments is widespread and affects many users of online services, we have evidence to suggest that this measure may be particularly useful for certain users. Members of a group with protected characteristics can be particularly targeted with illegal content through comments on their content. For instance, England football players Marcus Rashford, Jadon Sancho, and Bukayo Saka were targeted with racist abuse online in the aftermath of the Euro 2020 final. Some of this abuse was sent through comments on Instagram and X (formerly known as Twitter).[1755] Similarly, a three-year global study on gender-based online violence against women journalists, reported by UNESCO, found that nearly three-quarters of a sample group said they had experienced online violence in the course of their work. The study found that comment control functionalities helped victims feel safer online.[1756]

Unsocial Spaces [accessed 22 October 2024]; Phillips, J G., Diesfeld, K. and Mann, L., 2019. Instances of online suicide, the law and potential solutions, 26 (3). [accessed 28 October 2024].

[1752] See chapter 8 of this Volume: 'U2U settings, functionalities, and user support' for more information.

[1753] Between January and March 2023, YouTube removed more than 853 million comments from videos for violating its Community Guidelines. Of these, more than 44 million were for harassment or bullying, and over 87 million were due to child safety concerns. Source: YouTube, 2022. YouTube Community Guidelines enforcement - Google Transparency Report. [Accessed 22 October 2024].

[1754] Ofcom research found that of the respondents who had experienced hateful, offensive, or discriminatory conduct online in October 2021 to May 2022, 47% came across it in comments on or replies to a post, article, or video. Source: Ofcom, 2022. Online Experiences Tracker Data tables waves 1 and 2. [accessed 22 October 2024].

[1755] Landler, M., 2021. After Defeat, England's Black Soccer Players Face a Racist Outburst, New York Times, 12 July. [accessed 22 October 2024].

[1756] International Centre for Journalists, 2022, The Chilling: A global study of online violence against women journalists. [accessed 17 October 2023].

12.79 Users whose content is encountered by many others, such as high-profile figures, may also particularly benefit from this measure. There are several cases where high-profile figures have publicly announced their decision to disable the comments section on their uploads to social media after receiving abusive comments from other users.[1757] Additionally, when high-profile figures may receive a volume of abusive comments on their content, protecting themselves by blocking all non-connected users may not be desirable given their public status and muting individual accounts may not be feasible given the 'cascade effect'.[1758] Therefore, the ability to turn off comments on specific content provides an additional way for users to protect both themselves and others who can see their content from illegal harms.

12.80 As explained in the 'Summary of stakeholder feedback' section, Snap said that its current comment control functionality – involving manual approval of comments received – provides greater protection for users than our proposed measure.[1759] While service providers are free to meet their duties through alternative means, we consider our proposed measure will offer effective protection against the relevant harms on the range of services to which it applies. Given the severity and prevalence of the harms in question, this will deliver significant benefits. In giving users the opportunity to disable comments at the point of posting content, our measure will prevent other users from commenting on their content at all. For users that choose this, it will avoid any risk of them encountering illegal content through comments on their content.

12.81 We therefore consider that this measure will be effective in reducing the harm to which users are exposed.

### Making the measures known to users

12.82 Requiring service providers to make these controls known to users will make users more likely to use them. This will help protect users from illegal harms for the reasons described earlier in this section.

## Costs and risks

12.83 For each measure, we discuss the direct costs to service providers from its implementation and its potential indirect costs, along with stakeholder feedback on both types of cost.

### Measure on user blocking and muting

#### Direct costs of implementation

12.84 As set out in the 'Summary of stakeholder feedback' section, paragraph 12.28, we received feedback from industry stakeholders about the potential complexity and resource burden of these two measures. They combined their feedback on both measures and expressed concern that these measures are not straightforward to implement, that there could be technical or financial limitations to their implementation, and that there could be potential

[1757] Galluci, N., 2019, Taylor Swift says turning off Instagram comments does wonders for self-esteem, Mashable, 6 March 2019 [accessed 17 October 2024]. Miller, B., 2024, Selena Gomez reveals why she disabled Instagram comments, Independent, 20 May 2024 [accessed 19 November 2024].

[1758] Research conducted by Professor Matthew Williams found that hateful comments exposed to other users with corresponding thoughts or views may encourage them to do the same, resulting in a "cascade effect" of abuse against the victim. Source: Williams, M., 2019. Hatred Behind the Screens: A Report on the Rise of Online Hate Speech, (p.26) [accessed 22 October 2024].

[1759] Snap response to November 2023 Consultation, p.23.

disruption to user experience.[1760] Some respondents noted that this would pose a particular challenge for service providers with more limited resources, giving examples of smaller or non-commercial service providers.[1761] [✂].[1762]

12.85 In our November 2023 Consultation, we set out the estimated one-off direct and ongoing costs that would be incurred by relevant service providers not currently implementing the recommendation. We estimated that there would be an initial engineering effort of 20 to 150 days of software engineering time to implement the measure (with potentially an equal amount of time input from staff in professional occupations), resulting in an estimated one-off direct cost in the region of £9,000 to £140,000. In addition, we assumed an annual maintenance cost of 25% of the one-off cost, which is approximately £2,500 to £35,000 per year. We recognise that, in practice, many service providers already have the measure (or parts of it) in place. The costs for these services may therefore be lower than we have set out.

12.86 We also noted that, in some circumstances, there may be some overlapping costs with the implementation of our measure which stops users from sending messages to non-connected users in chapter 8 of this Volume: 'U2U settings, functionalities, and user support'.[1763]

12.87 We estimated a wide cost range, which reflects that there is likely to be considerable variation across service providers. We received no responses on alternative assumptions for the specific direct costs that we estimated to implement this measure, and we are unaware of any evidence to suggest there are more appropriate alternative assumptions about the amount of time it would take to implement this measure. We have therefore not changed the quantified cost assumptions that led to the estimates.

12.88 We consider it technically feasible to implement this measure, even for more complex services. However, we acknowledge that direct costs are likely to vary significantly across providers. Costs will be influenced by the design of the service and will depend on factors including the complexity of the provider's systems and the service's functionalities, the nature of how users interact on a service, and the extent of organisational overheads required to implement changes. Costs are likely to increase for larger services which tend to be more complex. Where services have many functionalities relevant to this measure, and for technical or organisational reasons there are few overlapping costs in the implementation of this measure between functionalities, the direct costs of the measure could be relatively high, at the top end of our quantified range, or even beyond in some cases.

12.89 In light of the above analysis, our estimate of the labour input to implement the measure is unchanged. We have updated the associated cost estimates in line with the latest wage data released by the Office for National Statistics ('ONS') and now estimate the measure will have a one-off direct cost in the region of £10,000 to £150,000, and an annual

[1760] [✂]; Global Network Initiative response to November 2023 Consultation, p.17; Mid Size Platform Group response to November 2023 Consultation, p.11.

[1761] Global Network Initiative response to November 2023 Consultation, p.17; [✂].

[1762] [✂]; [✂].

[1763] This measure reduces the risk of harm from grooming. For a detailed explanation, see chapter 8 of this Volume: 'U2U settings, functionalities, and user support'.

maintenance cost of 25% of the one-off cost, which is approximately £2,500 to £37,500 per year.[1764]

12.90 We recognise that service providers may need to reallocate resources or acquire additional resources to implement this measure. We have concluded that this measure is effective and proportionate for the services we recommend it for that have relevant risks and functionalities, both in and of itself, and as part of the overall package of codes measures, as explained in chapter 13 of this Volume: 'Combined impact assessment'. As noted above, the Act provides flexibility for those who do not wish to adopt Codes measures, allowing them to adopt alternative measures to meet their duties.

### Potential indirect costs

12.91 We recognise the potential for this measure to have an indirect impact on service providers and service users. Global blocking of all non-connected users could fundamentally alter the community or usage of a site, and users may be less likely to interact with or see content created by other unknown users if they choose to use this tool. This could reduce user engagement and use of a service, which could lead to reduced revenue as an indirect result of the measure. Interaction between user accounts differs across U2U services, according to the functionalities that are employed. This means considerable variation of this impact across different services.

12.92 However, any impact of the measure on engagement and usage rates is difficult to predict, and it is not necessarily always the case that use of a service and revenue will fall. While the overall effect on engagement may be negative for some users, there may be a countervailing positive impact to other users of a service which could help mitigate some of this impact. For example, users may disengage with services where they encounter illegal content. If users feel safer online due to the availability of blocking and muting tools, they may engage more with a service, albeit potentially with fewer users. Without such measures, some users may leave a service entirely.

12.93 The current availability and use of blocking and muting controls across different types of services (including individual blocking and global blocking for some functionalities) suggests that these controls add value for many services and users.[1765] [1766] The tools empower users to change their own experience of a service. As use of the feature is optional, it is reasonable to assume that the benefits that a user accrues from any instance of using blocking or muting tools must exceed the drawbacks to them personally, or they would not choose to use it.

[1764] We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by the ONS. However, since our cost estimates are rounded, the estimates have changed unevenly when using the updated wage assumptions. We received some feedback on the general cost assumptions (such as salary assumptions) that are fed into these costs. We consider that feedback in Annex 5.

[1765] Services including X, Facebook, Instagram, TikTok, LinkedIn, Snapchat, YouTube, Medium and Tumblr offer user blocking and/or muting tools. Source: Pen America, Online harassment Field Manual: Blocking, Muting and Restricting. [accessed 17 October 2024].

[1766] While less widely offered, there is evidence that some services currently provide users with the option to globally block all non-connected accounts for some functionalities. For example, Instagram allows for pre-emptive blocking of new accounts of the user of a blocked account, and also enables users to block all direct messages. Source: Meta, 2021, Introducing new tools to protect our community from abuse. [accessed 17 October 2024]. Discord allows users to block direct messages from users on a server that are not on their friends list. Source: Discord, 2022, Blocking & Privacy Settings. [accessed 17 October 2024].

12.94 As set out in the ‘Summary of stakeholder feedback’ section, several stakeholders said that this measure is not appropriate for all services, including Match Group and Wikimedia Foundation, who said that all or part of this measure presented a risk to its service (or service types) because it could impact how the service works and its useability, representing an indirect cost of the measure on the provider.[1767]

12.95 While dating services (and other similar types of services like friendship network services) are used to connect to new people, a global blocking function would enable users to continue to interact with their existing connections once these have been made, while not expanding their set of connections. We are aware of dating services which already offer a global blocking function or similar tool that allows users to prevent unconnected users from encountering their profile, but enables them to continue interacting with connected users.[1768] We agree that, if a user turned on global blocking before having any connections on this kind of service, this would negatively impact their experience of the service by preventing all U2U interactions. We consider that there would therefore be no incentive for a user to do so. As part of this measure, we are recommending that services inform users about the effect of using this tool, to enable them to make an informed choice about using it. We consider this mitigates the risk of it adversely impacting the experience of users on the service. We therefore do not consider that this measure will fundamentally undermine the useability of dating services or adversely impact their business model.

12.96 Regarding Wikimedia Foundation’s concerns, we understand the importance of editors being free to have open debates on Wikipedia Talk pages for the effective running of the service. We appreciate that in certain circumstances these tools could be misused by users to stifle debate, which in turn could impact the accuracy of Wikipedia articles. Recognising that it will not be known whether Wikipedia is in scope of these measures until it has been risk assessed, we have reviewed the functionality of Wikipedia Talk Pages in considering Wikimedia Foundation’s response to our consultation. We note that, while they are an important tool for editors, if user A were to block user B to prevent user B from replying to user A’s comments as part of a legitimate debate, user B could open another topic thread on the same subject to enable debate if needed. Additionally, users can edit Wikipedia articles without referring to their edit in the corresponding Talk Page.[1769] Overall, we recognise that the measure may lead to some degree of friction or inefficiency as part of the services’ functions in certain cases, but we still consider this proportionate in such cases so that users have control tools to mitigate the risk of harms targeted by this measure.

12.97 As discussed in paragraph 12.68, services may also wish to go further than the recommended measure by offering users additional controls to block specific user

[1767] Match Group response to November 2023 Consultation, p.17. We note that Match Group made the same point in response to the May 2024 Consultation, p.6; Wikimedia Foundation response to November 2023 Consultation, pp.33-34; Dwyer, D., response to November 2023 Illegal Harms Consultation, p.9.

[1768] Dating services which offer a form of blocking all unconnected users include: Hinge allowing a ‘pause’ which prevents a user from being shown to new people. Source: Hinge, Can I temporarily pause my Hinge account? [accessed 17 October 2024]. Bumble allows users to enable ‘snooze’ mode which hides users profiles from potential matches. Source: Bumble, What is snooze? [accessed 17 October 2024]. Tinder enables users to ‘disable Discovery’ where a user’s profile won’t be shown as a recommendation to new people. Source: Tinder, Hide your Tinder profile. [accessed 17 October 2024].

[1769] If a user were to edit an article, make note of this on a Talk Page, and then block some or all other users from seeing their comment in the Talk Page, this would provide no greater risk to the service than a user making an edit without referring to it in the Talk Page.

interaction on a service, for example blocking where direct messaging is prevented, but where users can still see content posted by the blocked user. This would have the dual benefit of giving users even greater control over how they protect themselves from harm and helping services reduce the impact on their business models, should some users decide they do not need to make use of the full extent of the blocking functionality as recommended in these measures.

12.98 On balance, we do not consider there to be any excessive or costly indirect impacts on services of different types due to this measure. We expect that service providers will be able to manage any indirect impact that this measure may have on their services.

### Measure on disabling comments

#### Direct costs of implementation

12.99 In our November 2023 Consultation, we set out the estimated costs of implementing this measure. A provider that does not currently offer the option to disable comments would incur one-off costs to make system changes and update the user interface. We estimated there would be an initial engineering effort of five to 50 days of software engineering time (with potentially an equal amount of time input from staff in professional occupations), resulting in a one-off direct cost of £2,000 to £50,000. In addition to one-off direct costs, we estimated ongoing maintenance costs of 25% of the one-off costs, which is approximately £500 to £12,500 per year. However, given this measure entails adapting the functioning of an existing comments function, rather than building a new functionality outright, we would generally expect costs to be lower than those for introducing the blocking and muting features as outlined in paragraphs 12.67-12.72.

12.100 As set out in the 'Summary of stakeholder feedback' section, we received feedback from industry stakeholders about the potential complexity and resource burden of this measure.[1770] This feedback was given about both measures, and we address it in paragraphs 12.84-12.90 above, where we discuss the direct costs of the blocking and muting measure. Our response to this stakeholder feedback regarding the disabling comments measure is the same.

12.101 As we are unaware of any specific evidence to suggest that our cost assumptions and estimates are inappropriate, our view remains unchanged from the November 2023 Consultation. Our estimate of the labour input to implement the measure is unchanged. We have updated the associated cost estimates in line with the latest wage date released by ONS. We now estimate the measure will have an estimated one-off direct cost in the region of £2,500 to £50,000, and an annual maintenance cost of 25% of the one-off cost, which is approximately £600 to £12,500 per year.[1771]

12.102 As discussed in paragraph 12.90, while we recognise that service providers may need to reallocate resources or acquire additional resources to implement this measure, we consider that it is effective and proportionate.

[1770] [✂]; Global Network Initiative response to November 2023 Consultation, p.17; Mid Size Platform Group response to November 2023 Consultation, p.11.

[1771] See Annex 5 for details of our cost assumptions. We have updated the estimates since the November 2023 Consultation in line with the latest wage data released by ONS. However, since our cost estimates are rounded off, the estimates have changed unevenly when using the updated wage assumptions. We received some feedback on the general cost assumptions (such as salary assumptions) that are fed into these costs. We consider this feedback in Annex 5.

### Potential indirect costs

12.103 We recognise that potential indirect costs could arise, which are difficult to estimate. If users were to disable comments on a widespread basis, this could impact the ability of users to interact with content. Over time, this could lead to lower engagement with content on a service (including with non-harmful content), or even to users leaving the service, with potential for a consequential reduction in revenue for providers.

12.104 However, giving users the ability to disable comments may deliver some counterbalancing indirect benefits to services. We expect service providers to benefit from retention of users who might otherwise leave the service because of an inability to control comments. While the overall impact of these contrasting effects is difficult to measure, various providers have already implemented this measure or similar measures to give users greater control of comment functionality.[1772] This indicates that service providers are aware that these functionalities can add value to the user experience by allowing comments to be disabled in some circumstances.

12.105 We also recognised that giving users the ability to control comments under content could result in a negative impact by preventing other users from replying to posted content. This may be particularly relevant where the content is defamatory or fraudulent. For example, if a user disabled comments on a piece of fraudulent content, other users would not have the ability to comment and warn others of this. This could result in instances of users being defrauded, which might not have occurred if comments were allowed. We consider that this risk is mitigated by the reporting and complaints measures, which enable users to report illegal content, and a dedicated reporting channel for fraud for services at risk of this illegal harm.[1773]

12.106 As set out in the 'Summary of stakeholder feedback' section, several stakeholders said that this measure is not appropriate for all services, including Reddit and Wikimedia Foundation who said that all or part of this measure presented a risk to their service (or service type) because it could impact service functionality and useability for their users, representing an indirect cost of the measure on the provider.[1774] We have carefully considered these concerns and reviewed the functionality and existing practices on the types of service about which stakeholders have raised concerns.

12.107 We understand that there are services where discussion in comment threads is a core feature of the service. However, we do not consider that this measure poses a fundamental

[1772] Several U2U services have implemented this type of measure. Instagram enables users to disable all comments or block certain users from commenting. Source: Meta, 2021, Introducing new tools to protect our community from abuse. [accessed 17 October 2024]; Facebook allows users to choose who can comment on uploaded posts, giving users the choice between 'everyone', 'people you follow', 'your followers' or 'people you follow and your followers'. Source: Meta, Facebook Help Centre, Commenting. [accessed 17 October 2024]; TikTok allows users to disable comments on their videos, as well as setting rules around who can comment based on their connection. Settings for users under 16 are set to 'friends only' for comments by default. Source: TikTok, Commenting. [accessed 17 October 2024]; YouTube gives users the option to disable comments on videos at any point after the video has been uploaded, as well as blocking certain accounts from commenting. It also allows for comment disabling on livestreams. Source: Sprout Social, 2022, YouTube Comments: A Complete Guide, [accessed 17 October 2024].

[1773] Reporting and complaints Codes ICU D1 and ICU D14. For a more detailed explanation see chapter 6 of this Volume: 'Reporting and complaints'.

[1774] Reddit response to November 2023 Illegal Harms Consultation, pp.9-10 and 23; Wikimedia Foundation response to November 2023 Consultation, pp.33-34; Dwyer, D., response to November 2023 Consultation, p.9.

risk to the useability of such services. As users go to forums for the purpose of discussion, we consider that there is little incentive for users to post on discussion forums and then disable comments for reasons other than harm prevention. If a user chose to open a discussion or topic thread but chose to turn off comments for the sole purpose of stifling debate (either from the outset, or after certain opinions began to be shared), we note that other users could start a new discussion to allow users to debate the topic. We also note that some discussion-focussed services already offer the functionality to disable comments to sub-groups of users such as volunteer moderators, which suggest that service providers see value in having some form of the measure and indirect costs may not be significant.

12.108 As with the blocking measure, we acknowledge that this measure could impact Wikipedia editors' ability to debate article content on Talk Pages, particularly if some users decided to use the tool to stifle debate rather than to protect themselves from harm. However, we note that if a user were to open a topic thread on a Talk Page and then immediately close it for comment, other editors could open another topic thread on the same subject to enable debate if needed. We therefore do not consider that this measure presents a material risk to editors' ability to ensure the accuracy of Wikipedia articles, though we do acknowledge that by creating the potential for the duplication of topic discussions it could create some friction on how this service is designed to function. We consider any such friction to be proportionate given the benefits of this measure.

12.109 On balance, we do not consider there to be any excessive or costly indirect impacts on services of different types due to this measure. We expect that service providers will be able to manage any indirect impact that this measure could have on their services.

#### Costs and risks of making these measures known to users

12.110 In our November 2023 Consultation, we asked for feedback on whether the measures should include provisions regarding how controls are made known to users. Having considered this feedback, we have added a provision which recommends that providers offer users information regarding these measures. This additional provision will lead to some direct costs for services.

12.111 We recommend that in-scope service providers make these user controls known to users, including setting out the effect of using these tools, and that this information is easy to find and comprehensible. We have not been prescriptive about how service providers should make these measures known to users, in order to give them flexibility in implementing this aspect of the measure. The information could be displayed on an information page, or in the place(s) on the service where they can be used as part of the user journey, for instance through interstitials or banners, though we are not specifically recommending an approach. We have estimated the direct costs of this measure would take between half a working day and 10 working days of software engineering time, with potentially an equal amount of time input from professional occupation staff. Using our standard assumptions, we expect that the one-off direct cost to be in the region of £200 to £10,000, with annual maintenance costs at 25% of this being around £50 to £2,500 per year.[1775] This cost burden on services is likely to be mitigated to some extent by the flexibility the measure allows regarding how these controls are made known to users.

[1775] See Annex 5 for details of our cost assumptions.

12.112 In our May 2024 Consultation, we estimated the costs which service providers in scope of proposed measure PCU E1 about ‘provision of age-appropriate user support materials for children’ would incur to implement that measure.[1776] Service providers in scope of that measure would be very unlikely to incur any additional costs for making these controls known to users, as implementing measure PCU E1 would likely be sufficient to make these measures known to users.[1777]

## Rights impact

### Freedom of expression and freedom of association

#### Measure on user account blocking and muting

12.113 As explained in ‘Introduction, our duties, and navigating the Statement', as well as in chapter 14 of this Volume: ‘Statutory tests’, Article 10 of the ECHR sets out 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. We must not interfere with this right unless satisfied that it is prescribed by law, corresponds to a pressing social need, and is proportionate to the legitimate aim pursued. Article 11 of the ECHR contains a similar right to freedom of association.

12.114 Google argued that the blocking functionality raises freedom of expression considerations.[1778] The rights to freedom of expression and freedom of association include the right to receive and impart information, and to associate with others, but do not include a right to compel others to listen to you or associate with you when they do not wish to. Affected users would not be prevented from imparting information by means of the service beyond the user that has chosen to block or mute them. We therefore do not consider this measure to interfere with the free expression or association of users. Users choosing to block or mute other users are exercising their own rights to freedom of expression and freedom of association by limiting the information they impart and the people they associate with.

#### Measure on disabling comments

12.115 While not citing freedom of expression directly, both Reddit and Wikimedia Foundation expressed concerns about the impact that comment disabling could have on conversations facilitated on their services (see paragraph 12.28).[1779] We acknowledge that if a user chooses to switch off comments on their own uploaded content, this removes an interface through which other users may receive and impart information and ideas. However, this would be a choice made solely by the user concerned and would have no impact on the right of other users to express themselves freely on the service concerned in other ways.

12.116 If they do not have the ability to disable comments, some users may be discouraged from posting at all due to the risk of harmful illegal content. This consideration is particularly acute for certain people, such as public figures, those that are the targets of hate campaigns or rumours, or those who are members of a group with protected characteristics.

[1776] For further information, see “Protection of Children Code of Practice for user-to-user services”, Ofcom, 2024, proposal PCU E1, pp.37-38

[1777] “Consultation: Protecting children from harms online”, Ofcom,2024, Volume 5, Chapter 21, p.360.

[1778] Google response to November 2023 Consultation, p.65.

[1779] Reddit response to November 2023 Consultation, pp.23-24; Wikimedia Foundation response to November 2023 Consultation, p.34.

12.117 We therefore do not consider that the measure would amount to an infringement of any user's right to freedom of expression, and could in fact promote self-expression by removing the risks for vulnerable users associated with expressing themselves online.

12.118 Giving users the ability to disable comments under content could impact users' right to reply in the case of derogatory or defamatory claims about them in another user's content. However, if implemented as intended this measure would not prevent users responding with their own content and linking to original content.

### Privacy

12.119 We do not consider these measures would interfere with users' rights to privacy, and indeed might have positive benefits for users' rights to privacy in that it would give them additional options for deciding how to share their personal information and content online.

### Data protection

12.120 We acknowledge that implementing the functionalities to give users the option to block or mute user accounts, or to disable comments on their own posts, will likely involve the processing of users' personal data. However, we consider that the amount of additional processing by the provider to implement these functionalities is likely to be minimal. In any case, service providers will need to comply with data protection law when carrying out any such processing.

### Making the measures known to users

12.121 There are no rights considerations relating to the recommendation for services to make these measures known to users.

## Who these measures apply to

12.122 In our November 2023 Consultation, we proposed recommending our blocking and muting measure (ICU J1) for large U2U services that:

- have identified as medium or high risk for any of the specified harms (coercive or controlling behaviour; harassment, stalking, threats and abuse; hate; grooming; encouraging or assisting suicide);[1780]
- have user profiles; and
- have at least one of the following functionalities: user connections; posting content; user communication (including but not limited to direct messaging and commenting on content).[1781]

12.123 In the same consultation we proposed recommending our measure on disabling comments (ICU J2) for large U2U services that:

[1780] The measure as consulted on in the November 2023 Consultation applied to services which are at medium or high risk of 'Encouraging or assisting suicide' and 'Encouraging or assisting serious self-harm'. We have now taken 'Encouraging or assisting self-harm' out of scope as we have made sure the harms groupings only include priority offences, consistent with Parliament's decision that they should be a priority.

[1781] As set out in the 'Our decision' section, we have clarified the scope of the global blocking measure by stating that it only applies to services that have user connection functionality.

- have identified as medium or high risk for any of the specified harms (harassment, stalking, threats and abuse; hate; grooming; or encouraging or assisting suicide);[1782] and
- have the functionality of commenting on content.[1783]

12.124 Following the consultation, we have decided to proceed with the original approach we proposed for applying this measure. We have concluded that our approach is proportionate considering the scale and severity of the relevant harms online, our analysis of the effectiveness of the measure, the costs to service providers of implementing it, and its limited impact on user rights.

12.125 We received feedback from stakeholders on the issue of who these measures apply to that was relevant to both measures.

### Size of service

12.126 In our November 2023 Consultation we consulted on the proposal to apply these measures only to providers of large services. We received a range of stakeholder feedback on this point.

12.127 As set out in the 'Summary of stakeholder feedback' section, several stakeholders argued that the measures should extend beyond providers of large services to providers of all sizes of service with relevant functionalities and risks, to offer more protection to users.

12.128 BT Group and Snap said that these measures represent basic or fundamental design decisions, with Snap emphasising that it is important to design these features at an early stage. Snap further argued that recommending these measures exclusively to providers of large services risked giving a competitive advantage to providers of smaller services.[1784]

12.129 The Board of Deputies of British Jews, and UK Safer Internet Centre argued that these measures should apply to all services regardless of size.[1785] Age Verification Providers Association and VerifyMy argued that the measures should extend to child users on all sizes of service with relevant risks.[1786]

12.130 Regarding the potential to impact competition, based on our analysis of impacts on services and the fact that a number of large services already have similar measures in place, we consider it highly unlikely that these two measures will have a significant impact on overall competitive dynamics between large and smaller services. For more information about how we have taken into account the size of a service when considering proportionality in accordance with our specific duties under Schedule 4 of the Act, see chapter 13 of this

[1782] The measure as consulted on in the November 2023 Consultation applied to services which are at medium or high risk of 'Encouraging or assisting suicide' and 'Encouraging or assisting serious self-harm'. We have now taken 'Encouraging or assisting self-harm' out of scope as we have made sure the harms groupings only include priority offences, consistent with Parliament's decision that they should be a priority.

[1783] We define "commenting on content" as a functionality that allows users to reply to content or post content in response to another piece of content and is visually accessible directly from the original content without navigating away from that content.

[1784] BT Group response to November 2023 Consultation, p.2; Snap response to November 2023 Consultation, p.23.

[1785] Board of Deputies of British Jews response to November 2023 Consultation, p.3; UK Safer Internet Centre response to November 2023 Consultation, p.12. We note that UK Safer Internet Centre made the same point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.37.

[1786] Age Verification Providers Association response to November 2023 Consultation, p.3; VerifyMy response to November 2023 Consultation, p.13.

Volume: 'Combined impact assessment'.[1787] As we note in that section, in the event that smaller services grow in size and reach seven million monthly UK users, they would become large services and would be in scope of measures recommended for such services.

12.131 In contrast some stakeholders argued that the potentially resource-intensive nature of these measures could result in a high level of cost and complexity,[1788] including for small services, and said they would have concerns if we recommended these measures for them.[1789] As set out in the 'Costs and risks' subsection, we have estimated a large range of expected direct costs for implementation. While costs will tend to increase for larger services due to their increased complexity, there are many factors which will determine the costs for a specific service, and it is possible that costs could be high even for smaller services.

12.132 Given the significant benefits of protecting users from the relevant types of illegal content, we are confirming our decision to apply these measures to large services. We consider that it is proportionate to recommend these measures for providers of large services that are at risk of the specified types of content and that have relevant functionalities, and this decision is supported by our evidence and stakeholder feedback at consultation.

12.133 In our May 2024 Consultation, we proposed recommending two equivalent measures for U2U services of any size (not only large services) that are likely to be accessed by children, have the same relevant functionalities, and have risk of certain types of content harmful to children. We have also received stakeholder feedback on who these measures apply to in response to that consultation. We acknowledge the difference between the size of services for which we proposed recommending these measures in our November 2023 Consultation and for the equivalent measures in the May 2024 Consultation. We also note the arguments that some respondents have made for extending the illegal harms measures to smaller services with relevant risks. Here we are confirming our decision to apply these measures to large services in the Illegal Content Codes. In the upcoming Protection of Children Statement, we will both determine who these measures apply to in our Children's Safety Codes and consider whether there is a case for extending the scope of the blocking and muting measure and disabling comments measure in the Illegal Content Codes to capture smaller services. This will allow us to take a holistic view of how we address the risks of these functionalities on smaller services across our two sets of Codes, whilst still bringing important protections on large services into force at the earliest opportunity.

### Service type

12.134 GNI said that the potentially resource-intensive nature of these measures would be challenging for non-commercial services.[1790] We do recognise that not-for-profit service providers may have fewer resources than commercial services. However, we maintain that all service providers need to consider the level of risk of their service, and well-intentioned services can still carry a risk of illegal content which must be addressed.

12.135 As discussed in this chapter, a number of service providers, Match Group, [✂], Booking.com, Reddit and Wikipedia, argued that the were not proportionate for their type of service. Examples of the types of services raised by stakeholders included dating services,

[1787] Volume 2, chapter 13: 'Combined impact assessment'.
[1788] Mid Size Platform Group response to November 2023 Consultation, p.11.
[1789] Global Network Initiative response to November 2023 Consultation, p.17; [✂].
[1790] Global Network Initiative response to November 2023 Consultation, p.17.

discussion forums, and video sharing platforms (VSPs).[1791] We have primarily addressed this feedback in earlier sections because it also relates to other themes such as the perceived benefits to users, effectiveness of the measure, or potential direct or indirect costs to services for these service types.

12.136 These measures are not targeted to apply to or exclude specific types of service. During our consideration of stakeholder feedback, we considered the possibility of only applying the measures to specific types of U2U service or excluding certain types of U2U service. However, at this time, we do not consider that would be appropriate for a number of reasons.

- As outlined in the 'Benefits and effectiveness' section, we consider these measures have the potential to combat different kinds of illegal harm across a broad range of service types and across different functionalities of services. These measures are risk based, and so services are only in scope where they have identified at least one medium or high risk of a relevant illegal harm. If a service is not risky, the measure does not apply to them.
- Furthermore, in the 'Costs and risks' section, we explained that we do not consider the direct and indirect costs to be disproportionately high for services in scope.
- More generally, in most cases we do not set Codes measures based on the types of service as this would add complexity to their application. It would also not be a practical approach in most cases, as many services would not fit neatly into a typology, potentially causing uncertainty for service providers.

12.137 We have therefore decided to apply these measures to all U2U service providers within scope, regardless of the type of U2U service.

### Functionality

#### Scope of our measures overall

12.138 We received feedback from Google that the functionality of posting content should not be sufficient (in addition to other requirements) to bring service providers in scope of the blocking and muting measure. It is concerned that the same provisions will apply equally to services with full social media functionalities as those with minimal social functionality.[1792]

12.139 While we acknowledge that posting content is a very common feature on U2U services (meaning that this will bring many services within scope), the measure will only apply to services that also have user profiles, and where their illegal harms risk assessment has identified a medium or high risk of the relevant illegal harms. Furthermore, as set out in the 'Benefits and effectiveness' section, posting content can cause users to encounter illegal content. As set out in the 'Costs and risks' section, where services have fewer types of functionalities where users interact with each other (or with each other's content), they are likely to incur lower costs of implementation because they will have fewer service features to change in order to implement the measures.

[1791] Booking.com response to November 2023 Consultation, p.21; [✂]; [✂]; Match Group response to the May 2024 Consultation, p.6; Match Group response to November 2023 Consultation, p.17; Reddit response to November 2023 Illegal Harms Consultation, pp.9-10 and 23; Wikimedia Foundation response to November 2023 Illegal Harms Consultation, pp.33-34.

[1792] Google response to November 2023 Consultation, p.59; we note that Google made the same point in response to the May 2024 Consultation, p.40.

### Scope of global blocking provisions

12.140 As set out above, these measures apply to services with specific functionalities. As outlined in the 'Our decision' section, we have clarified the scope of the global blocking aspect of the measure by stating that it only applies to services that have user connection functionality. This part of the measure is designed to give users the option to interact (for instance, by viewing content and exchanging direct messages) only with users with whom they have chosen to connect through the service's user connection functionality (for instance, Facebook's 'Add friend' functionality).

12.141 Without this additional wording in the measure, providers of services without user connection functionality may have sought to implement this measure by categorising all other user accounts as 'unconnected' to the blocking user. This could have led to blocking users largely experiencing the service as if there were no other users on the service, which was not the intention of this measure.

## Conclusion

12.142 Our analysis suggests that the benefits associated with these measures are material. While the measures will result in some direct costs for service providers, we do not consider these to be disproportionately high when set in the context of the benefits the measures will deliver. This is particularly the case given that we are focusing the measures in question on large services, for whom we anticipate the costs will normally be manageable. In addition, we have concluded that any indirect costs of the measures will be proportionate and that the measures will not have an undue adverse impact on fundamental rights – instead, they could benefit the fundamental rights of victims and survivors. We have therefore decided to proceed with recommending the measures, with an additional provision that recommends providers make these measures known to users and clarifying the scope of the global blocking part of the measure.

12.143 The measures are in the Illegal Content Codes of Practice for U2U services on CSEA and other duties. They are referred to in these Codes as ICU J1 and ICU J2.

# Measure on notable user and monetised labelling schemes

12.144 In our November 2023 Consultation, we proposed that services operating "notable user" or "monetised" schemes should:

- provide profile information to help users to understand what profile labelling means in practice; and
- have clear internal policies for these schemes and consistently apply them.

12.145 We proposed this measure to apply to all providers of large services that are assessed as medium or high risk of fraud or foreign interference, and which have a notable user and/or a monetised scheme. We proposed this measure over alternative options of (1) relying on the user empowerment and user identity verification duties for Category 1 services in the Act and (2) proposing that large online services should establish and maintain a notable user scheme that meets certain criteria.

12.146 The measure proposed was designed to mitigate the risks from impersonation, a tactic used by perpetrators engaging in forms of fraud and foreign interference. We considered that well-run notable user and monetised schemes can play a valuable role in enabling users to make informed decisions about the content they choose to interact with by helping them to identify potentially illegal content. On the other hand, poorly operated and communicated schemes may introduce more risks than benefits for users who place trust in them.

12.147 A 'notable user scheme' is a scheme under which a service provider labels a user's profile to indicate to other users that they are a 'notable user'.[1793] These users might include, but are not necessarily limited to, politicians, celebrities, influencers, company executives, journalists, government departments, non-governmental organisations, financial institutions, media outlets, and companies. The label indicating that a user is notable (for example a 'tick' symbol) may appear on that user's profile and/or any content they publish.

12.148 A 'monetised scheme' is one under which a provider labels the user profile of a user who has made a payment to the provider of the service or some other person. Such schemes may be open to all users and payment may be regular or one-off. Users participating in the scheme may benefit from access to additional features on the service. The label indicating that a user is participating in a monetised scheme may appear on that user's profile and/or any content they publish.

12.149 In the Codes measure and below, we use the phrase 'relevant schemes' to refer to notable user schemes and monetised schemes together. We note that stakeholders sometimes refer to these schemes as 'verification schemes', whether or not they actually carry out any steps to verify the identity of users. To avoid conflating these schemes with identity verification, we do not refer to them as verification schemes.

12.150 We refer to a user whose profile is labelled under a relevant scheme as a 'relevant user'. We use the term 'labelling' to refer to the symbol and associated words added to a user's profile under a relevant scheme.

## Summary of stakeholder feedback[1794]

12.151 Several stakeholders expressed support for the measure and did not suggest any changes to it.[1795]

[1793] A 'user profile' is a functionality, associated with a user account, that represents a collection of information shared by a user which may be viewed by other users of the service. This can include information such as username, biography, profile picture, etc., as well as user-generated content generated, shared or uploaded by the user using the related account.

[1794] Note this list in not exhaustive, and further responses can be found in Annex 1.

[1795] ACT The App Association response to November 2023 Consultation, p.17; Betting and Gaming Council response to November 2023 Consultation, p.12; Centre for Competition Policy response to November 2023 Illegal Harms Consultation, p.17; LinkedIn response to November 2023 Consultation, p.15; NSPCC response to November 2023 Consultation, p.40; OnlyFans response to November 2023 Consultation, p.9; Segregated Payments Ltd response to November 2023 Consultation, p.13; Stop Scams UK response to November 2023 Consultation, p.16; Welsh Government response to November 2023 Consultation, p.4.

12.152 Some stakeholders, while offering support, suggested amendments or noted risks linked to the measure.[1796] Others, as set out below, suggested amendments to the measure without offering support. These risks and proposed amendments related to:

- proportionality of our measure;
- importance of distinguishing between fraud and foreign interference;
- information provided to users;
- restrictions when labelling profiles of commercial entities;
- verifying the identity of relevant users;
- recommending providers establish notable user schemes;
- risks of notable user and monetised schemes; and
- feedback on who this measure applies to.

12.153 We lay these out in the following paragraphs.

## Proportionality of our measure

12.154 Google raised several concerns regarding the proportionality of elements linked to the internal policy expectation:

- Google suggested removing the recommendation for providers to establish criteria and thresholds for the labelling of profiles.[1797] It stated that this does not "enable flexibility for platforms who may have different verification/labelling schemes for a wide range of users who hold a wide variety of different positions and/or roles", adding that it is "unclear how platforms would meaningfully meet this requirement to mitigate the intended risks".
- It also questioned the proportionality of the proposed expectation that service providers set out safeguards to ensure that user profile information is not modified without the provider reviewing and consenting to that change. It noted the example of Channel 4 which alters its YouTube bio to share links to currently available popular shows.[1798]
- It also raised concerns about our expectation that service providers set out in their internal policies how they will treat relevant users and the content they post on the service. It said it was "generally unclear of Ofcom's intention for including these provisions" and recommended we delete them, arguing that "the Codes and Act already

[1796] 5Rights Foundation response to November 2023 Consultation, pp.31-32; Age Verification Providers Association response to November 2023 Consultation, p.3; Association of British Insurers (ABI) response to November 2023 Illegal Harms Consultation, pp.2-3;; CELE response to November 2023 Consultation, p.13; [✂]; Clean Up the Internet response to November 2023 Consultation, pp.4-5; Innovate Finance response to November 2023 Illegal Harms Consultation, p.16; Local Government Association response to November 2023 Consultation, p.13; Match Group response to November 2023 Consultation, pp.17-18; Mencap response to November 2023 Consultation, p.14; Meta response to November 2023 Consultation, annex, pp.15-16; Mid Size Platform Group response to November 2023 Consultation, p.11; Monzo response to November 2023 Illegal Harms Consultation, pp.19-20; OneID response to November 2023 Consultation, p.3; Snap response to November 2023 Consultation, pp.23-24; South East Fermanagh Foundation response to November 2023 Illegal Harms Consultation, pp.17-18; VerifyMy response to November 2023 Consultation, p.13; Which? response to November 2023 Illegal Harms Consultation, p.15.

[1797] Google response to November 2023 Consultation, p.60.

[1798] Google response to November 2023 Consultation, pp.60-61.

cover most of these issues, and it seems unnecessary to duplicate these provisions (and almost impossible for platforms to implement parallel functionalities in this area if that was the intention)".[1799]

12.155 We address these concerns in the 'How this measure works' section.

## Importance of distinguishing between fraud and foreign interference

12.156 The Cyber Threats Research Centre stated that "Foreign interference/political influence campaigns should not be conflated with other types of online harms like fraud", and also that the "way the Foreign Influence Offense (FIO) has been framed appears abstracted as though it has not yet occurred".[1800]

12.157 We address this concern in the 'Benefits and effectiveness' section.

## Information provided to users

12.158 Many respondents highlighted the need for users to be able to understand what profile labelling means and why it is done.[1801] Several respondents also suggested that we factor in the additional accessibility needs of some groups to ensure sufficient user understanding. Mencap made this argument regarding users with a learning disability, while 5Rights Foundation and UK Safer Internet Centre (UKSIC) suggested children should be able to understand any information provided.[1802]

12.159 We address these comments in the 'Benefits and effectiveness' section (specifically 'Effectiveness of public-facing information about relevant schemes' at paragraph 12.203).

## Restrictions when labelling profiles of commercial entities

12.160 One stakeholder suggested there should be basic restrictions to prevent criminals obtaining a labelled user profile, [✂].[1803] The Association of British Insurers ('ABI') suggested that any verification of corporate entities should go further than checking they exist and are registered at Companies House.[1804]

12.161 We address these comments in the 'Benefits and effectiveness' (specifically 'Arguments supporting further expectations on providers to check credentials' at paragraph 12.215).

## Verifying the identity of relevant users

12.162 We received various responses on, or linked to, the topic of identity verification ('IDV'). These points were often linked to evidence of harm from anonymous or fake user profiles, which identity verification could help to prevent. We address in this chapter the responses that supported identity verification specifically in the context of relevant schemes.

[1799] Google response to November 2023 Consultation, p.61.

[1800] Cyber Threats Research Centre Swansea University response to November 2023 Illegal Harms Consultation, p.2.

[1801] Betting and Gaming Council response to November 2023 Consultation, p.12; Cifas response to November 2023 Illegal Harms Consultation, p.18; Electronic Frontier Foundation response to November 2023 Consultation, p. 18; Institute for Strategic Dialogue response to November 2023 Consultation, p.13; INVIVIA response to November 2023 Consultation, pp.23-25; Match Group response to November 2023 Consultation, p.18; [✂]; South East Fermanagh Foundation response to November 2023 Consultation, p.17; UK Safer Internet Centre response to November 2023 Consultation, pp. 15-16.

[1802] 5Rights Foundation response to November 2023 Consultation, p.32; Mencap response to November 2023 Consultation, p.14; UK Safer Internet Centre response to November 2023 Consultation, p.16.

[1803] [✂].

[1804] ABI response to November 2023 Consultation, p.3.

Feedback about identity verification more widely is addressed in chapter 11 of this Volume: 'User access'.[1805]

12.163 Several responses suggested providers should have more rigorous processes to verify the identity of users with labelled profiles. Logically and Snap made this point regarding the identity of those labelled as "notable users".[1806] The Local Government Association, Cifas, and Lloyds Banking Group raised this point in relation to monetised schemes due to the perceived credibility that perpetrators can obtain by paying for their profile to be labelled.[1807] In making this argument, the latter two respondents supported a role for the Government's Digital ID Trust Framework.[1808]

12.164 Clean Up the Internet and NCA suggested that there should be processes to verify the identity of users with labelled profiles generally, with the latter highlighting risks of recovery fraud and noting that identifiers (such as phone numbers and email addresses) could be used to identify perpetrators.[1809] Which? suggested that Ofcom's work on user empowerment could inform future amendments to this measure in relation to user verification.[1810] [1811]

12.165 We address this feedback in the 'Benefits and effectiveness' section (specifically 'Arguments supporting further expectations on providers to check credentials' at paragraph 12.215).

### Recommending providers establish notable user schemes

12.166 The Association of British Insurers and Logically said that we should recommend providers who do not already have notable user schemes to establish them (Option B in the November 2023 Consultation).[1812]

12.167 We address these comments in the 'Costs and risks' section.

### Risks of notable user and monetised schemes

12.168 Several respondents highlighted concerns about the risks of these schemes. A common concern was that monetised schemes could give perpetrators an opportunity to create a false impression of legitimacy that could be used to carry out fraud or spread disinformation.[1813] For example, Cifas and Clean Up The Internet argued X's introduction of a monetised scheme "created an illusion of authenticity and trust, which strengthen the

[1805] See Volume 2: chapter 11: User access.

[1806] Logically response to November 2023 Consultation, p.10; Snap response to November 2023 Consultation, p.24.

[1807] Cifas response to November 2023 Consultation, p.18; Lloyds Banking Group response to November 2023 Consultation, p.10; Local Government Association response to November 2023 Consultation, p.13.

[1808] Cifas response to November 2023 Consultation, p.18; Lloyds Banking Group response to November 2023 Consultation, p.10.

[1809] Clean Up the Internet response to November 2023 Consultation, p.4-5; NCA response to November 2023 Consultation, pp.58-59.

[1810] Which? response to November 2023 Consultation, p.15.

[1811] This falls under Ofcom's work on additional duties for categorised services. More information is available here: https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/roadmap-to-regulation/ .

[1812] Association of British Insurers response to November 2023 Consultation, pp.2-3; Logically response to November 2023 Consultation, p.10.

[1813] Are, C. response to November 2023 Consultation, p.17; Cifas response to November 2023 Consultation, p.18; Clean Up the Internet response to November 2023 Consultation, p.4; Lloyds Banking Group response to November 2023 Illegal Harms Consultation, p.10; South East Fermanagh Foundation response to November 2023 Consultation, p.17; The Cyber Helpline response to November 2023 Consultation, p.19; Which? response to November 2023 Consultation, p.15.

scammers position" and meant "bad actors could easily obtain the credibility of a "blue tick" without going through any form of meaningful verification" respectively.[1814] The Cyber Threats Research Centre, the NCA and Meta raised risks that perpetrators may seek to compromise accounts with labelled profiles and game requirements to attain a label.[1815] The Institute for Strategic Dialogue argued that "overly broad categories" for labelling could lead to "obscuring variations in the trustworthiness of different sources".[1816] Protection Group International said that some users will attempt to get their profile labelled to avoid certain repercussions, noting that some service providers do not take action against users participating in relevant schemes even though they are offending in some way.[1817] Which? agreed with the measure's intention to expect providers to have "good design practices" for relevant schemes.[1818]

12.169 We address some of these concerns and comments in the 'Benefits and effectiveness' section (specifically 'Consistently applied internal policies on the operation of relevant schemes'). We also address some of these points in the 'Costs and risks' section.

## Feedback on who the measure applies to

12.170 A number of respondents stated that the measure should apply to all services in scope of the Act, or be extended to smaller services that are also medium or high risk for fraud or the foreign interference offence.[1819] For example, Which? suggested that the risk of poorly designed schemes necessitates the measure applying regardless of size, while UKSIC suggested that broader application could specifically be useful against harms such as grooming, harassment, and cyberflashing.[1820]

12.171 Wikimedia Foundation argued that it would be negatively affected, perceiving the measure as being designed for social media services.[1821]

12.172 Google, GNI, Match Group, Mid Size Platform Group and techUK made the same point on this measure as on the blocking, muting and disabling comments measures. They highlighted the variety of service types in scope, noting the importance of considering the proportionality of the measure (see 'Summary of stakeholder feedback' under the measures on blocking and muting, and disabling comments for details).[1822]

[1814] Cifas response to November 2023 Consultation, p.18; Clean Up the Internet response to November 2023 Consultation, p.4.

[1815] Cyber Threats Research Centre Swansea University response to November 2023 Consultation, p.2; Meta response to November 2023 Consultation, annex, pp.15-16; NCA response to November 2023 Consultation, pp.58-59.

[1816] Institute for Strategic Dialogue response to November 2023 Consultation, p.13.

[1817] Protection Group International response to November 2023 Consultation, p.11.

[1818] Which? response to November 2023 Consultation, p.15.

[1819] Age Verification Providers Association response to November 2023 Consultation, p.3; Cifas response to November 2023 Consultation, p.18; Logically response to November 2023 Illegal Harms Consultation, p.10; OneID response to November 2023 Consultation, p.3; Online Safety Act Network response to November 2023 Illegal Harms Consultation p.81; Snap response to November 2023 Consultation, p.23; UK Safer Internet Centre response to November 2023 Consultation, pp.15-16; Which? response to November 2023 Consultation, p.15.

[1820] UKSIC response to November 2023 Consultation, pp.12-13; Which? response to November 2023 Consultation, p.15.

[1821] Wikimedia Foundation response to November 2023 Consultation, pp.33-34.

[1822] Google response to November 2023 Consultation, pp.58 and 60; Global Network Initiative response to November 2023 Consultation, p.17; Match Group response to November 2023 Consultation, p.18; Mid Size Platform Group response to November 2023 Consultation, p.11; techUK response to November 2023 Consultation, p.27.

12.173 We address these comments in the ‘Who this measure applies to’ section (specifically ‘Service size’ and ‘Service type’ at paragraphs 12.237-12.240).

## Our decision

12.174 We have decided to broadly confirm the measure we proposed in the November 2023 Consultation. We have made three minor changes in response to the feedback set out in the previous section:

- Rather than expecting providers to review changes to profile information whenever they are made, our recommendation is now that that providers’ internal policies should set out safeguards to ensure that a user profile is not modified to suggest the user account is operated by or on behalf of anyone other than the relevant user. We explain our reasons for making this change in paragraph 12.184.
- We have amended our expectation about service providers’ internal policies describing whether relevant users are treated differently on the service. We have clarified that this information is to be recorded only where relevant users are treated differently from other users. The measure sets out that this information should include how relevant users are treated differently in relation to recommender systems, content moderation and account security. This list of topics has been reduced in our final measure. We have removed: ‘content curation’, ‘reporting and complaints’, ‘quality assurance’ and ‘fact-checking’. We explain our reasons for making this change in paragraph 12.185.
- We have specified that providers should ensure user-facing descriptions of relevant schemes are clear and accessible. Rather than refer to measure ICU G3[1823], we have set out equivalent expectations to reflect that the information provided may not always be incorporated into a provider’s Terms of Service. We explain our reasons for making this change in paragraph 12.180.

12.175 The full text of the measure can be found in our Illegal Content Codes of Practice for other duties and is referred to as ICU J3.

## Our reasoning

### How this measure works

12.176 This measure recommends that, where providers operate ‘notable user’ or ‘monetised’ schemes (defined in paragraphs 12.147 and 12.148), they should:

- provide information on the profile of a relevant user to indicate why and under which scheme the profile is labelled (referred to from hereon in as ‘profile information’) and a user-facing description of the scheme so users can understand what profile labelling means in practice; and
- have, and consistently apply, internal policies for operating relevant schemes.

[1823] This terms of service and publicly available statement measure recommends all U2U and search service providers ensure that relevant provisions included in terms and statements regarding the protection of individuals from illegal content are clear and accessible, including being comprehensible based on the likely reading age of the youngest individual permitted to use the service without the consent of a parent or guardian. See chapter 10 in this Volume: ‘Terms of service and publicly available statements’.

12.177 In this section, we set out our expectations of providers to act in accordance with the measure in our Codes. We firstly explain what should be covered in the profile information and user-facing descriptions of relevant schemes, including how the latter should be made accessible. Secondly, we review our expectations of what consistently applied internal policies should consider. We also set out what needs to be considered in the design, communication, and review of internal policies. Finally, we explain the changes we have made compared to the initial proposal about what should be addressed by internal policies.

### Public-facing information about notable user and monetised schemes

12.178 To act in accordance with the measure, profiles labelled by providers under a relevant scheme should include information on why the profile has been labelled. Where the provider operates more than one scheme, profiles should indicate under which scheme the profile is labelled.

12.179 Separately, providers should publish a user-facing written description which explains how and why user profiles are labelled under a relevant scheme, and how and why a label may be removed. The description should be consistent with what is set out in the provider's internal policies (described in paragraphs 12.183-12.185), though we recognise it may be less detailed than those policies to avoid providing users with information enabling them to game a scheme.

12.180 We expect this description to be clear and accessible. We have amended the measure in the Codes to explain how this should be achieved, rather than cross-referring to measure ICU G3. We consider this approach to be clearer for providers, and to take into account that providers can decide whether to include this information in their Terms of Service or elsewhere.

### Consistently applied internal policies on the operation of relevant schemes

12.181 We expect internal policies to be designed to reduce any risk of harm to users from fraud and/or the foreign interference offence, as identified by a service's risk assessment. The policies should be communicated to relevant staff, including through regular training, as well as being regularly reviewed and updated to ensure they remain fit for purpose. When reviewing their policies, providers should take into account at least one of the following:

- user feedback and reporting;
- user experience testing; and
- engagement with persons with relevant expertise.

12.182 The measure sets out that internal policies should also include specific information. As summarised in paragraph 12.154, Google raised several concerns regarding elements of this part of the measure. We note throughout this section our consideration of these points and where they have led to changes to the proposal set out in our November 2023 Consultation.

12.183 **Criteria for labelling:** For all schemes, a provider's policies should set out the process for considering and thresholds for deciding whether to label a user profile or to remove a label. For notable user schemes, the policies should also set out how the provider will satisfy itself that the account is operated by or on behalf of the person by whom or on whose behalf it is held out as being operated. The policies should also set out how the provider will satisfy itself that, if a notable user is presented as holding a particular position or role, they actually hold that position or role.

- As noted, Google recommended deleting this expectation. It argued that it "does not enable flexibility for platforms who may have different verification/labelling schemes for a wide range of users who hold a wide variety of different positions and/or roles".[1824] We do not accept this argument, as the measure does not interfere with providers' discretion to determine on what basis they label the profiles of users and how they determine whether the user in question holds the role they say they do under a notable user scheme. Instead, the measure seeks to ensure that the processes the provider follows are recorded in its internal policy documents, including the criteria and thresholds it applies for labelling profiles under any relevant schemes it operates. The design of the measure maintains flexibility for service providers to operate multiple kinds of labelling schemes, but as explained above in paragraphs 12.178-12.179, recommends that the differences between the schemes are clearly displayed and explained to users.

12.184 **Ensuring labelling remains appropriate:** A provider's policies should set out safeguards on how it will ensure that the profile information (such as username and 'bio' text) on a profile labelled under a notable user scheme is not modified to suggest the user account is operated by or on behalf of someone else. The policies should also set out the frequency with which the provider will conduct reviews to confirm whether the user profiles of relevant users should continue to be labelled, and the circumstances in which it will carry out such reviews.

- We originally proposed that a service provider's policies should set out safeguards to ensure profile information is not modified without the provider reviewing and consenting to that change. In its response, Google stated that this was "far too wide an intrusion on users' ability to update their own channels/content". It gave the example of Channel 4 changing their 'About' information to promote popular shows.[1825] We considered this feedback and concluded that a provider approving each change to profile information could be disproportionate. Nevertheless, we consider that the measure should still address the risk of notable users' profile information being changed to carry out harm through impersonation. We have therefore amended how the measure describes the expected safeguards. It now recommends that a provider's internal policies should set out safeguards to ensure that the profile information of relevant users is not modified to suggest the account is being operated by someone else. We consider that this change will allow providers sufficient flexibility to develop and deploy processes that are appropriate for their service and the schemes they operate, while effectively mitigating the risk of users changing their profile information to carry out harm through impersonation.

12.185 **If and how relevant users are treated differently on the service:** If providers treat relevant users and the content they post on the service differently from other users, then the policies should set this out and explain how they do so. Policies should include a description of if and how relevant users are treated differently from other users in relation to recommender systems, content moderation, and account security. As noted in paragraph 12.150 we use the term 'relevant users' to mean those whose profiles are labelled under a relevant scheme.

[1824] Google response to November 2023 Consultation, p.60.

[1825] Google response to November 2023 Consultation, p.60.

- In response to our proposal that policies should include a description of if, and how, relevant users are treated differently, Google said it was unclear why we had included this expectation as proposed. It argued that it duplicated other parts of the Codes and the Act. We agreed in part that this expectation could be clearer. However, we did not accept Google's view that this expectation duplicates others in the Code and Act, given the specific focus on the operation of notable user and monetised schemes. We have therefore amended the measure to clarify that the internal policy should only cover this issue if relevant users are treated differently to other users and, if so, should include how they are treated in relation to content moderation, recommender systems and account security. We have removed "reporting and complaints" as we consider it could unnecessarily duplicate the expectation set by including "content moderation". We have removed "quality assurance" and "fact-checking" to ensure it is clear to providers what we expect under this measure. We have also deleted "content curation" as we consider it could unnecessarily duplicate the expectation set by including "recommender systems".

## Benefits and effectiveness

### Benefits

#### Risk of harm

12.186 This measure aims to address the risk of harm from fraud and foreign interference posed to users from the impersonation of high-profile individuals and organisations. For instance, users could fall victim to impersonation fraud or believe disinformation disseminated by a hostile foreign state actor after trusting content which has been posted by accounts with labelled profiles. This could happen if users do not understand what labelling on another user's profile conveys, or if a labelled profile is misused by a perpetrator.

12.187 This risk of harm from this is considerable. The National Fraud Intelligence Bureau recorded £2.35 billion as lost to all types of fraud (including but not limited to impersonation fraud) in 2020 to 2021.[1826] [1827] Beyond money lost, the 2023 UK Government Fraud Strategy estimated that the total economic and social cost of fraud to individuals between 2019 and 2020 was £6.8 billion.[1828] The chapter of the Register of Risks ('Register') titled 'Fraud and financial services' gives further detail of the harm caused by fraud.[1829]

12.188 Impersonation of UK media outlets and think-tanks is a tactic used by those engaging in forms of foreign interference. Where users come to see these sources as credible, their engagement and sharing of content increases the risk of foreign interference. The chapter of the Register 'Foreign interference' gives further detail of the harm caused by foreign interference.[1830]

12.189 This measure seeks to address these risks by highlighting the importance of users knowing why a profile has been labelled and to understand what that label signifies. It seeks to balance this harm mitigation with maintaining flexibility for service providers to be able to make their own decisions about users' experience of the service. The measure rests on the

[1826] National Fraud Intelligence Bureau, 2021. Fraud Crime Trends. [accessed 31 October 2024].

[1827] Impersonation fraud is where perpetrators trick users into engaging with fake online accounts imitating those of high-profile individuals, UK Government departments such as HM Revenue and Customs or the Department for Work and Pensions, and financial institutions such as banks, for example.

[1828] Home Office, 2023. Fraud Strategy. [accessed 17 October 2024].

[1829] Register of Risks chapter titled 'Fraud and financial services'.

[1830] Register of Risks chapter titled 'Foreign interference'.

premise that service providers partly choose to operate notable user and monetised schemes to build a sense of credibility for their services and for relevant users. We therefore consider it reasonable to expect providers to take the steps in this measure to help mitigate risks associated with relevant schemes.

12.190 In terms of these risks, the Cyber Threats Research Centre said that "Foreign interference/political influence campaigns should not be conflated with other types of online harms like fraud" and argued that the way we had framed the foreign interference offence "appears abstracted as though it has not yet occurred".[1831]

12.191 The Act sets out a list of priority offences, and treats fraud offences and foreign interference offences separately. Our measure takes the same approach – a large service is in scope of the measure if it is at medium to high risk of one or both harms.[1832] The reason our measure seeks to mitigate the risk of both fraud and foreign interference is because poorly operated user labelling schemes may enable users to carry out impersonation, and our evidence suggests this is a tactic used to commit both these kinds of illegal harm. However, we recognise that the way the harms manifest may differ. Furthermore, while disinformation and misinformation spread by hostile foreign states is a long-standing concern and form of activity, the specific offence of foreign interference referred to in the Act is new.

12.192 Protection Group International noted that some service providers do not take action against users participating in relevant schemes even though they are offending in some way.[1833] Where providers treat relevant users and the content they post differently to other users, this measure recommends this be explained in the provider's internal policies. However, the duty on providers to operate their services using systems and processes designed to swiftly take down illegal content when they become aware of it applies regardless of the user who has posted that content.

12.193 Overall, we consider there continues to be a case for a measure that addresses the risk of harm from fraud and foreign interference using impersonation as a tactic.

### Risks presented by profile labelling

12.194 In regard to the risk presented by profile labelling, our own research shows that users take profile labelling into account when considering the authenticity and accuracy of online content, and thus the importance of the factors contained in our measure.

12.195 Our research has examined the attitudes of UK users towards profile labels. Ofcom's Media Literacy Trackers examine the experience of UK internet users, including their online skills and confidence. In the Adults Media Literacy Tracker, respondents were shown a social media post and asked to judge whether it appeared to be genuine and explain why they came to their conclusion. Nearly seven in 10 (67%) participants said they felt that the social media post was genuine. The most important features of the post that led them to make that decision were the name of the page from which the content was posted and the blue

[1831] Cyber Threats Research Centre Swansea University response to November 2023 Consultation, p.2.

[1832] As set out in the Risk Assessment Guidance, service providers should assess their services for the combined risk of fraud and financial services offences. See: Risk Assessment Guidance and Risk Profiles, Part 1, Table 1: List of illegal content to assess.

[1833] Protection Group International response to November 2023 Consultation, p.11.

'verification' tick.[1834] Separately, in the Children's Media Literacy Tracker, among the 95% of respondents aged 12 to 17 who correctly identified an NHS Instagram post as genuine, three in 10 (27%) said the presence of a 'verified' tick was an indicator of credibility.[1835] Furthermore, in a poll conducted for us by YouGov, nearly three in ten (28%) UK internet users aged 16 or over claimed that they used 'verification' labels when deciding to follow or interact with an account on social media.[1836] Therefore, our initial evidence base suggests that profile labelling is an important factor taken into account by users when deciding whether to engage with content or assessing if it appears to be genuine.

12.196 We also carried out a literature review to establish whether there is a broader evidence base for the claim that providers' schemes are important for establishing user trust and content credibility and have a meaningful effect on user behaviour. The literature specifically linking credibility to labelling or account authenticity is limited.[1837] However, the results we found generally supported the fact that users are aware of profile labelling schemes and that account credibility is a factor in increasing user trust.[1838]

12.197 We concluded from this literature review that labelling and account authenticity are likely to increase trust and credibility. However, they interact with a variety of other factors such as knowledge of the scheme and celebrity or congruity (e.g., when influencers promote material that does not 'fit' with user perceptions of the account holder).

### Industry practice

12.198 Our observations of the risks that arise from current industry practice, and how providers have responded to these, indicate that providers being transparent about profile labelling can be beneficial for protecting users.

[1834] This research involved showing social media users a real social media post and asking them if they thought it was genuine or not, and to give their reasons for doing so. Of the 67% of adult social media users who thought a Citizens Advice post was genuine, 45% identified the verification tick as amongst their reasons for making this judgement. Source: Ofcom, 2024. Adults' Media Use and Attitudes report 2024, pp.23-33. [accessed 6 September 2024].

[1835] Respondents aged 12 to 17 who go online were shown a real Instagram post and asked whether they thought it was genuine or not, and to give their reasons for their opinion. Source: Ofcom, 2024. Children and Parents: Media Use and Attitudes, pp.36-37. [accessed 6 September 2024].

[1836] Respondents were asked "when using social media platforms, how often, if at all, do you look out for these kinds of labels (e.g., a tick on a profile) when deciding to follow or interact with an account?". Nearly three in ten respondents (28%) claimed they "always" (2%), "often" (7%), or "sometimes" (19%) used verification labels when deciding to follow or interact with an account on social media. A further fifth (22%) said they used these labels "rarely", suggesting that these respondents may find verification labels helpful in certain contexts or situations. Source: Ofcom, 2023. Verification Schemes to Label Accounts poll. [accessed 21 September 2023].

[1837] This is in part due to academic interest in how other factors interact with schemes (e.g. perceptions of celebrity or trust) and because the majority of studies are focused on X, given the public knowledge of the first so-called blue check scheme and the length of time since it was launched it in 2009.

[1838] Morris, M. R., Counts, S., Roseway, A., Hoff, A., Schwarz, J., 2012. Tweeting is Believing?: Understanding Microblog Credibility Perceptions, *Proceedings of the 15th ACM Conference on Computer Supported Cooperative Work*, 441–450. [accessed 23 August 2023]; Kapitan, S., Silvera, D., 2016. From digital media influencers to celebrity endorsers: attributions drive endorser effectiveness, *Marketing Letters*, 27 (3), 553-567. [accessed 23 August 2023]; Edgerly, S., Vraga, E. K., 2019. The Blue Check of Credibility: Does Account Verification Matter When Evaluating News on Twitter, *Cyberpsychology, Behaviour, and Social Networking*, 22 (4), 283-287. [accessed 23 August 2023]; Vaidya, T., Votipka, D., Mazurek, M. L., Sherr, M., 2019. Does Being Verified Make You More Credible? Account Verification's Effect on Tweet Credibility, *CHI Conference on Human Factors in Computing Systems Proceedings*. [accessed 23 August 2023]; Taylor, S. J., Muchnik, L., Kumar, M., & Aral, S., 2023. Identity effects in social media, *Nature Human Behaviour*, 7 (1), 27-37. [accessed 23 August 2023].

12.199 During 2022 and 2023, X and Meta introduced monetised schemes having had a precedent for labelling notable profiles.[1839] [1840] Media sources analysed the impacts at the time, commenting that accounts belonging to the original and new monetised schemes appeared to be displayed with the same label.[1841] [1842] [1843] This raised questions about whether users would be able to clearly distinguish between profiles that were verified under the different relevant criteria for notability or for a paid subscription.

12.200 Since the November 2023 Consultation, we have continued to review publicly available information about notable user and monetised schemes on a range of services including Facebook and Instagram, X, TikTok, LinkedIn, Snapchat, YouTube, and Pinterest.[1844]

12.201 The continued risk of impersonation of individuals and abuse of labelled profiles is clear from several providers' information pages setting out what protections and checks they have in place:

- Pinterest notes in its information pages that all requirements for being a Verified Merchant must continue to be met, otherwise a user will be suspended from the scheme.[1845]
- Google provides a link to report impersonation of users or channels on YouTube.[1846]
- TikTok states that reasons for removing a 'verified badge' may include username change and account type change.[1847]

[1839] X launched a new subscription scheme in November 2022 and had a period of transition until April 2023. It includes different coloured checkmarks for different kinds of users, including for businesses and government or multi-lateral organisations. Source: X, 2024. About X Premium and About profile labels and checkmarks on X. [accessed 28 August 2024].

[1840] Meta developed a new paid-for verification scheme called Meta Verified, partly aimed at helping creators to establish an online presence, and launched it in the UK in May 2023. Source: Meta, 2023. Testing Meta Verified to help creators. [accessed 22 September 2023]; Meta, 2024. Stand out with Meta Verified. [accessed 28 August 2024].

[1841] inews reported that X explained to users that: "The blue checkmark may mean two different things: either that an account was verified under Twitter's previous verification criteria (active, notable, and authentic), or that the account has an active subscription to Twitter's new Twitter Blue subscription product…". Source: McCann, J., 2022. What do the Twitter blue, yellow and grey tick mean?, inews, 14 December. [accessed 22 September 2023].

[1842] The symbol on X meant two different things until April 2023 when X began to remove blue ticks from accounts that did not join the new subscription service. Source: Digital World, 2023. Twitter puts end to blue tick for users who don't pay, 21 April. [accessed 24 August 2023].

[1843] Tech Crunch reported that "there is no visual differentiation between a legacy verification badge and the new subscription badge for Meta Verified" accounts. TechCrunch, 2023. Meta's paid verification program is now available in the UK, 17 May 2023. [accessed 24 August 2023].

[1844] Google, 2024. Verification badges on channels. [accessed 28 August 2024]; LinkedIn, 2024. Identity verification via Persona. [accessed 28 August 2024]; Meta, 2024. Stand out with Meta Verified. [accessed 28 August 2024]; Meta, 2024. Understanding Verification on Facebook and Instagram [accessed 28 August 2024]; Pinterest, 2024. Apply to join the Verified Merchant Programme. [accessed 28 August 2024]; TikTok, 2024. Verified accounts on TikTok. [accessed 28 August 2024]; X, 2024. About X Premium. [accessed 28 August 2024]; Snap, no date. How to Verify Your Public Profile. [accessed 28 August 2024].

[1845] Verified Merchants are vetted brands who benefit from additional features promoting their products. Source: Pinterest, 2024. Apply to join the Verified Merchant Programme. [accessed 28 August 2024].

[1846] Google, 2024. Verification badges on channels. [accessed 28 August 2024].

[1847] TikTok, 2024. Verified accounts on TikTok [accessed 28 August 2024].

12.202 Furthermore, we continue to see examples of impersonation by those potentially engaging in relevant harms which pose a risk to users.[1848]

12.203 We therefore consider the evidence on how users interact with profile labelling demonstrates the need to address the risks that could materialise from poorly run notable or monetised schemes.

### Effectiveness

#### Effectiveness of public-facing information about relevant schemes

12.204 As set out in paragraphs 12.178 to 12.180 in the 'How this measure works' section, providers are expected to provide public information in the form of profile information and user-facing descriptions. We also specify that the user-facing descriptions about schemes need to be clear and accessible. As set out in the 'Summary of stakeholder feedback' section, paragraphs 12.158, some respondents highlighted the importance of bearing in mind the needs of children and people with learning disabilities. As set out in paragraph 12.168, some highlighted their concerns about perpetrators using monetised schemes to give a false impression of legitimacy.

12.205 Providing descriptions on user profiles in combination with a user-facing description of the scheme should help users to understand why a provider has labelled a user's profile, enabling them to take this context into consideration when making decisions about whether to engage with content that could cause them harm.[1849] We also consider that the expectations in the measure that a user-facing description be clear and accessible will mean service providers create descriptions that account for a range of user needs. We are not prescriptive about how this should be provided and therefore at this stage do not think it appropriate to specify particular ways in which it should be accessible to disabled people. We would expect approaches to accessibility to vary from service to service, subject to their services features and design and on that basis recommend providers are best placed to decide how to ensure information is accessible to disabled people. However, providers should consider their obligations under other relevant legislation (for example, the Equality Act 2010) and, where relevant, appropriate guidance.[1850]

12.206 We have observed considerable variation in the depth of explanations provided to users. We therefore consider it necessary to set baseline expectations for what should be included. Though we recognise that not all users will read the information provided in a user-facing description, it may be summarised and cascaded by other users who have an interest in analysing these features.

12.207 While we consider that providing this public information will help to protect users from the relevant harms wherever these schemes are in place, it will be particularly useful where providers operate multiple schemes aimed at different sections of their userbase or as part of monetisation or other strategies, due to the risk of confusion among users as to why a particular profile has been labelled. If users find it difficult to determine under what scheme

[1848] Desai, R. 2024. Gold Rush on the Dark Web: Threat Actors Target X (Twitter) Gold Accounts *CloudSEK Whitepapers and Reports.* [accessed 28 August 2024]; Which?, 2024. Travel scams: airline customers targeted by fake accounts on X. [accessed 28 August 2024].

[1849] Account labels are often a badge, tick, or checkmark symbol. This is generally prominent on a relevant user's profile and may also be visible where users see and engage with content posted by the relevant user. It is likely that these labels are the most widely seen information about relevant schemes.

[1850] For example, World Wide Web Consortium's (W3C) Web Content Accessibility Guidelines (WCAG).

a profile has been labelled, in a worst-case scenario that difficulty could be exploited by subscribers to a monetised scheme by impersonating an entity that could be considered notable.

12.208 For example, Martin Lewis, the Executive Chair of the UK's biggest consumer help site, stated in April 2023 that a profile with a verified 'X Premium' subscription checkmark was impersonating him to promote a cryptocurrency.[1851] This meant there was a clone that could mislead users about financial advice and abuse the trust that other users place in both Martin Lewis's reputation and the fact the profile was labelled. Media reporting indicates that other publicised instances of the apparent misuse of verification schemes have attracted attention and confusion.[1852] These examples highlight how users could place trust in a profile intending to deceive for fraudulent purposes or covert influence.

12.209 Overall, the public facing information we expect providers to have about relevant schemes is a crucial element that we consider will make this measure effective, especially given the evidence of how users interact with profile labelling.

### Consistently applied internal policies on the operation of relevant schemes

12.210 As described in paragraphs 12.181 to 12.185 in 'How this measure works', we expect providers to have internal policies setting out how they operate their relevant schemes. These should be designed to reduce the harm identified in the service's risk assessment and be implemented as set out in the policies.

12.211 Respondents offered support for this part of the measure, including Snap which noted the importance of service providers having a "rigorous set of processes" for labelling profiles of notable users, and Which? which agreed with the need for "good design practices".[1853] The risk of account compromise was highlighted by the Cyber Threats Research Centre, NCA and Meta, demonstrating the importance of providers having processes to ensure profile labelling remains appropriate.[1854]

12.212 There are several benefits flowing from service providers having internal policies that take into account the service's risk assessment and cover certain details about how schemes operate. We expect these to improve the effectiveness of the measure.

- **Criteria for labelling**: Clearly setting out thresholds and criteria for adding or removing profile labels should mean that providers act consistently when doing so. This should mean users can trust that a profile has been labelled for the reasons the provider sets out in its public facing information. There is an increased risk of harm if profiles are incorrectly labelled as notable. Therefore, there is a benefit in the policy setting out how a provider will satisfy itself that a notable user is who they say they are, as this should increase consistency in the application of these policies. This should lead in turn to users being able to trust that a notable user profile has been labelled for the reasons the provider sets out in its public facing information.

[1851] Tweet by Martin S Lewis of MoneySavingExpert on 3 April 2023. [accessed 22 September 2023].

[1852] Sardarizadeh, S., 2022. Twitter chaos after wave of blue tick impersonations. BBC News, 12 November. [accessed 24 August 2023].

[1853] Snap response to November 2023 Consultation, p.24; Which? response to November 2023 Consultation, p.15.

[1854] Cyber Threats Research Centre Swansea University response to November 2023 Consultation, p.2; Meta response to November 2023 Consultation, annex, pp.15-16; NCA response to November 2023 Consultation, pp.58-59.

- **Ensuring labelling remains appropriate**: We expect a provider's policies to include safeguards against notable users' profile information being changed to suggest that the account is operated by someone else. We consider the benefit of this to be twofold. Firstly, it should mitigate against the risk of a user deliberately seeking to take advantage of their labelled status to perpetrate harm against other users. Secondly, it should also reduce the risk of harm occurring where a relevant user's account is compromised by a perpetrator and the perpetrator seeks to change the profile information to deceive other users. In addition, a provider's policy addressing how and when profiles labelled under either type of relevant scheme are reviewed should help to ensure that the provider's criteria for labelling continue to be met.
- **If and how relevant users are treated differently**: Where providers treat relevant users differently to other users, this should be recorded in the policies, including in relation to content moderation, recommender systems and account security. We know providers may apply some processes differently to relevant users, such as in relation to recommender systems (for example, X prioritises replies from accounts using 'X premium') and account security (for example, TikTok's requirement for accounts to have 2-step verification in place before they can receive a 'verified badge').[1855] [1856] Where providers decide to treat relevant users differently to other users, the measure expects them to design this part of their policies in a way that reduces the risk of fraud and foreign interference as identified in the service's latest risk assessment. This should mitigate the risk of providers' policies, including in relation to recommender systems, content moderation and account security, exacerbating the risks caused by impersonation connected to relevant schemes.

12.213 If a service provider designs or adjusts a relevant scheme to reduce the risks of fraud and/or foreign interference as identified in the most recent risk assessment of its service, this should ensure that the scheme seeks to protect users from these harms. Communicating internal policies on relevant schemes to relevant staff should lead to more consistent application, reducing, for example, the risk of incorrect labelling of a user profile. It will also be beneficial for users if, when reviewing their policies, providers take into account one or more of (1) user feedback and reporting, (2) user experience testing and (3) engagement with persons with relevant expertise. This is because external feedback about the schemes should enable providers to take into account users' firsthand experience of how the schemes are working in practice, and to reflect industry best practice.

12.214 For these reasons, we consider that consistently applied internal policies about relevant schemes are beneficial and necessary for the measure to be effective.

### Arguments supporting further expectations on providers to check credentials

12.215 We set out our expectations of providers under 'How this measure works'. This measure seeks to benefit users by expecting providers to provide public facing information about schemes and consistently apply internal policies about how they operate. It does not specify the checks that providers should carry out when labelling users, instead allowing providers flexibility to operate their schemes in a way that is appropriate for their service. Below we respond to feedback suggesting the measure would be more effective if we

[1855] X, 2024. About X Premium. [accessed 28 August 2024].

[1856] TikTok, 2024. Verified accounts on TikTok. [accessed 28 August 2024].

recommended that providers check the profiles of commercial entities and verify the identity of relevant users.

12.216 As noted in paragraph 12.160 in the summary of feedback, stakeholders said the measure should set expectations regarding checks for commercial entities. [✂], while the ABI recommended that verification should go further than confirming existence and registration at Companies House.[1857] Relatedly, as summarised in paragraphs 12.163 to 12.164 of the summary of stakeholder feedback section, some stakeholders argued further that our measure would be more effective if it set expectations regarding verifying the identity of users.[1858]

12.217 Regarding the specific case of checking the credentials of commercial entities, while we recognise the potential benefit of recommending providers do this, we do not consider that introducing this change at the current time is critical to the measure being effective. However, we may consider introducing specific checks to this measure in future, subject to consideration of the value, feasibility and potential unintended consequences of doing so. We also note that the Illegal Content Judgements Guidance includes a reference to the FCA Warning List for the purpose of helping service providers identify illegal financial promotions when reviewing content.[1859]

12.218 More broadly on verification, we recognise the potential benefits of IDV in protecting users from fraud and foreign interference. However, we are not at this time proposing to add any measures recommending IDV to our Illegal Content Codes. This is because we will be considering IDV as we progress our work on the user identity verification duties for categorised services, which is in phase 3 of our work plan. Considering IDV at this point will enable us to take a holistic view on the issue. Feedback about identity verification more widely is addressed in chapter 11 of this Volume: 'User access'.[1860]

12.219 Based on our analysis, we consider that this measure will be effective in reducing the risk of users becoming a victim of fraud or foreign interference. Given the prevalence and impact of fraud and the potential detrimental impacts of foreign interference, we consider that the measure in question will deliver important benefits to users.

## Costs and risks

### Costs

12.220 Service providers that currently operate a relevant scheme that does not meet the recommendations of the measure will incur some costs.

12.221 We set out an analysis of potential costs for implementing this measure in the November 2023 Consultation, as follows in this section. We did not receive any feedback specifically relating to costs which led us to change our analysis.

[1857] Association of British Insurers response to November 2023 Consultation, p.3; [✂].

[1858] Cifas response to November 2023 Consultation, p.18; Clean Up the Internet response to November 2023 Consultation, p.4-5; Lloyds Banking Group response to November 2023 Consultation, p.10; Local Government Association response to November 2023 Consultation, p.13; Logically response to November 2023 Consultation, p.10; NCA response to November 2023 Consultation, pp.58-59; Snap response to November 2023 Consultation, p.24; Which? response to  November 2023 Consultation, p.15.

[1859] Ofcom's Illegal Content Judgements Guidance: chapter 6 'Fraud and financial offences'.

[1860] Volume 2: chapter 11: User access.

12.222 Costs could include developing or improving appropriate internal policies and processes and training staff to apply them. The policies would need to cover which users are eligible for labelling and how providers will apply them. As staff working on this feature would need to have a good understanding of these policies, additional training is likely to be needed.

12.223 Providers will need to consider whether their relevant schemes could be better designed to decrease the risk of harm to users (for example, to reduce the risk that users misunderstand the schemes). When changes are needed, providers will incur additional costs in redesigning their schemes. This could represent a substantial cost if a scheme needs to change materially. While the policies and training associated with this measure may be more than some providers currently have, we expect that all service providers currently running such schemes will already have some policies and training in place, meaning that it is unlikely they would need to redesign the scheme from scratch. This is likely to reduce the total additional cost.

12.224 Providers may need to improve transparency for users about what profile labelling means. They may incur design and engineering costs when making necessary system changes, such as making the description of the relevant scheme easily accessible to users.

12.225 For some users, including perpetrators, joining a monetised scheme has the appeal of attaining a label similar to those appearing on notable users' profiles. The measure recommends that providers take steps to ensure the differences between the schemes are clearly communicated to users. This could disincentivise some users from joining such schemes, potentially reducing revenue. However, there are other reasons why some users join monetised schemes. The monetised schemes we have studied state that there are other benefits, such as access to exclusive features. This suggests that ensuring users understand the difference between the schemes will have only a limited impact on take-up.

12.226 As set out in the 'Summary of stakeholder feedback', paragraph 12.165, some respondents said that we should recommend providers introduce 'notable user' schemes for their services where they do not have such schemes. This was because they considered such schemes would help mitigate the risks we outlined about impersonation. As we said in our November 2023 Consultation, this would be both materially more intrusive and considerably more costly than the measure we have decided to recommend. At this point, we do not have evidence that recommending service providers to introduce such schemes would bring sufficient benefits to users to make this proportionate.

12.227 Implementation of the measure is likely to vary between services. The scale of the cost would depend on the changes (if any) made by a provider to apply the measure. As costs are likely to vary significantly on a service-by-service basis, we have not been able to assess the level of costs in a detailed way. Some service providers may only need to make small adjustments to the operation and/or communication of their schemes, while others may need to make more significant changes.

### Risks

12.228 We agree with the risks highlighted by respondents about how monetised schemes operate, as mentioned in paragraph 12.168. Respondents who raised this included Cifas and Clean Up The Internet, who argued X's introduction of a monetised scheme "created an illusion of authenticity and trust, which strengthen the scammers position" and meant "bad actors could easily obtain the credibility of a "blue tick" without going through any form of meaningful verification" respectively.[1861] We acknowledge that providers in part rely on the perceived credibility of labelled profiles to promote participation in a relevant scheme. When designing this measure, we were conscious not to interfere with the commercial freedom of providers to operate a relevant scheme. However, we consider poorly run schemes to give rise to risks of fraud and foreign interference, and this is what the measure intends to tackle, including by expecting providers to be transparent with their users about how a relevant scheme works in practice and what it means where a profile is labelled.

12.229 We are aware of the risks of account takeover that can arise where schemes are in place, as noted by the Cyber Threats Research Centre, Meta and the NCA.[1862] The Codes expect that providers' internal policies should include how account security steps may differ for relevant users. The measure also recommends that providers apply internal policies which cover safeguards to ensure that a notable user's profile information is not changed to suggest the user account is operated by or on behalf of someone else, which could be attempted by a perpetrator. We have also captured this risk in the Register.[1863]

12.230 The Institute for Strategic Dialogue mentioned that if the categories for labelling are too broad then this may create the perception of all labelled profiles as being equally trustworthy.[1864] The measure is not prescriptive in setting criteria for labelling but through increased transparency and consistently applied internal policies about relevant schemes aims to enable users to better understand why a profile may be labelled.

12.231 We recognise that providers could choose to discontinue their relevant schemes if they find the costs of the measures excessive. We consider that providers are unlikely to remove a longstanding element of their services that some users appear to value. However, even if a provider were to cease operating a scheme, this would not necessarily harm users' interests overall. As set out in paragraph 12.186, some users may be at risk of fraud and foreign interference if they misunderstand the status of relevant users. The measure is designed to mitigate this – if a provider does not want to implement the measure, it would likely be better for such users if the scheme were discontinued.

## Rights impact

### Freedom of expression and freedom of association

12.232 We expect this measure to have little to no impact on users' freedom of expression, as we do not anticipate any of the recommendations set out in this measure to have any impact on users' ability to impart and receive information. In addition, all the relevant schemes currently operated by service providers within the scope of the measure are voluntary and a user's access to a service is not dependent on participating in a relevant scheme.

### Privacy

12.233 We recognise that, in formulating policies that are consistent with this measure, service providers may require users to provide personal information in order for their profile to be

[1861] Cifas response to November 2023 Consultation, p.18; Clean Up the Internet response to November 2023 Consultation, p.4.

[1862] Cyber Threats Research Centre Swansea University response to November 2023 Consultation, p.2; Meta response to November 2023 Consultation, annex, pp.15-16; NCA response to November 2023 Consultation, pp.58-59.

[1863] Register of Risks chapter titled 'Fraud and financial services'.

[1864] Institute for Strategic Dialogue response to November 2023 Consultation, p.13.

labelled under a relevant scheme. Given that participation in a relevant scheme is voluntary, we consider this to be proportionate.

### Data Protection

12.234 We acknowledge that service providers may choose to gather additional data from users in order to decide whether to label a user's profile under a relevant scheme. We recognise that providers may also retain certain data to implement any safeguards to prevent misuse of the account of a relevant user, and to carry out regular reviews of such accounts. This would likely involve retaining that data longer for relevant users than it is retained for other users. However, providers will have to comply with applicable privacy and data protection law when collecting and processing users' data, including ensuring that they do not collect any more data and do not retain such data for longer than is necessary for the purpose for which it is collected.

## Who this measure applies to

12.235 In the November 2023 Consultation, we set out that this measure should apply to providers of large services if they:

- identify a medium or high risk of fraud or foreign interference; and
- operate a notable user or monetised scheme.

12.236 We made these proposals based on the scale of the considerable potential harm posed by impersonation for the purposes of fraud or foreign interference. We considered the benefits of reducing these harms justify the costs associated with the measures, and that providers of large services are likely to have the resources to bear any costs of this measure.

### Service size

12.237 As set out in paragraph 12.170, several respondents argued that the measure should be extended to all providers in scope of the Act or be extended to smaller services that are also medium or high risk for fraud or the foreign interference offence. For example, Which? suggested that the risk of poorly designed schemes necessitates the measure applying regardless of size, while UKSIC suggested that broader application could specifically be useful against harms such as grooming, harassment, and cyberflashing.[1865]

12.238 We focussed on large services because their large user base is attractive for perpetrators seeking to reach large numbers of people at low cost and with minimal effort (as set out in the Register).[1866] A large user base may also make it more likely that the initial reach of fraudulent user generated content will be amplified to an even bigger potential audience via a higher volume of content reactions, posts, and reposts. As such, the benefits of applying the measure to large services are particularly material. Given this, and that we consider that large services are also likely to have the resources to bear any costs of this measure, we have decided to apply the measure to large services that operate a relevant scheme and are at medium or high risk of fraud or foreign interference.

12.239 The benefits of this measure are likely to be less on services with lower reach, and there is uncertainty on the precise implementation costs for such services. On the basis of the

[1865] UK Safer Internet Centre response to November 2023 Consultation, pp.15-16; Which? response to November 2023 Consultation, p.15.

[1866] Register of Risks chapters titled 'Fraud and financial services' and 'Foreign interference'.

evidence we currently have, it is not clear to us that the measure would be proportionate for smaller services. We have therefore not extended it to cover smaller services at this time.

### Service type

12.240 Wikimedia Foundation argued that it would be negatively affected by the measure because it believed it to have been designed for social media services.[1867] In response to both of our measures, respondents highlighted the variety of service types in scope, and noted that it is therefore important to consider the proportionality of the measure for these different service types.[1868] We acknowledge these concerns, but continue to believe that providers of large services operating relevant schemes (as defined in paragraphs 12.147 to 12.149) should apply this measure where they have a medium or high risk of fraud or foreign interference. We are aware there is variation among service types and in-scope schemes but are not aware of any reasons that the measure would be technically impossible or disproportionate for providers of certain types of service to implement.

## Conclusion

12.241 We have decided to include this measure in our Codes largely unchanged from the measure proposed in our November 2023 Consultation, with the three modifications referred to in paragraph 12.174 (under ‘Our decision’). We recommend that providers should have, and consistently apply, clear internal policies for operating notable user and monetised schemes on their services and should improve public transparency for users about what profile labelling means in practice. We have not sought to limit how providers develop various schemes on their service; instead, we have created recommendations that will ensure they do so with appropriate internal policies and transparency for users regarding how a relevant scheme operates.

12.242 Costs are likely to vary across services, and there is uncertainty on their precise level. However, we consider it proportionate to recommend this measure to large services that have assessed themselves as being at medium or high risk of fraud or foreign interference, and which operate notable user or monetised schemes. As we have explained, the measure will play an important role in mitigating risks associated with fraud and the foreign interference offence. The scale of the challenge posed by impersonation to commit fraud or foreign interference, and hence the potential benefits of the measure, are considerable when applied by large services and are likely to justify the costs associated with them. Large services are also likely to have the resources to bear any costs of this measure.

12.243 This measure will be contained in our Illegal Content Codes of Practice for other duties and is referred to as ICU J3.

[1867] Wikimedia Foundation response to November 2023 Consultation, pp.33-34.

[1868] Google response to November 2023 Consultation, pp.58 and 60; Global Network Initiative response to November 2023 Consultation, p.17; Match Group response to November 2023 Consultation, p.18; Mid Size Platform Group response to November 2023 Consultation, p.11; techUK response to November 2023 Consultation, p.27.

# 13. Combined Impact Assessment

### What is this chapter about?

In the preceding chapters in this volume we have assessed the impact of each of the measures we are recommending in this Statement. In this chapter, we assess the combined impact of the recommended measures as a package. Having considered the combined impact on different groups of services, we consider the package of measures to be proportionate.

## Introduction

13.1 The measures recommended in this Statement are designed to be the first step to protecting users from illegal harms online. The measures are set out in Table 1 and Table 2 in the 'Summary of our decisions'.

13.2 In the preceding chapters we have assessed the impacts of each of the measures and concluded they are proportionate. In this chapter, we consider the combined impact of the measures by looking at:

- Whether each measure has distinct benefits that contributes to how the overall package reduces risks of illegal 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. 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 regulatory burden is proportionate, particularly for smaller services. In assessing this, we recognise that even where the cost of individual measures may not be significant, the cost burden of the whole package may be significant for some service providers.

## Some stakeholders raised proportionality concerns

13.3 In our November 2023 Consultation, we considered the impact of the overall package of measures that we proposed for providers of different groups of services. We considered this for smaller services and large services with different risks. This reflects the variations in proposed measures based on the outcome of a service's risk assessment and its size. We provisionally concluded that the proposed package of measures recommended for each group of services was proportionate.

13.4 We received a range of responses on the proposed package of measures for different services:

- **Smaller service providers**: Many respondents agreed with our approach for providers of smaller services.[1869] Some argued that we should go even further and apply more measures to providers of smaller services that are risky for a single kind of illegal harm or low-risk for all kinds of harm.[1870] However, several respondents were concerned about the impact on small businesses as some smaller services might cease to operate.[1871] Some also cautioned that the package of measures could create significant barriers to entry for new services, or discourage businesses from moving to or developing in the UK. [1872]
- **Large service providers**: Many respondents considered the burden on large service providers to be reasonable, given their greater resources.[1873] Several stakeholders pointed out that the larger size of such services leads to a higher risk of causing harm and therefore requires more extensive measures.[1874] Other stakeholders suggested that our measures for providers of large services were not extensive enough.[1875] On the other hand, several stakeholders opposed the idea that size should influence the degree to which service providers are regulated, arguing that size is not always a good proxy for the risks a service entails,[1876] or its provider's financial capability.[1877] Some providers also argued that it was disproportionate to apply measures to providers of large low-risk services.[1878]

[1869] Local Government Association response to November 2023 Illegal Harms Consultation, p.16; [✂]; Nexus response to November 2023 Illegal Harms Consultation, p.21; OneID response to November 2023 Illegal Harms Consultation, p.6.
[1870] Board of Deputies of British Jews response to November 2023 Illegal Harms Consultation, p.6; Children's Commissioner response to November 2023 Illegal Harms Consultation, p.21; International Justice Mission response to November 2023 Illegal Harms Consultation, p.9.
[1871] Bolton, C. response to November 2023 Illegal Harms Consultation, p. 11; Dwyer, D. response to November 2023 Illegal Harms Consultation, p. 11; Name Withheld 3 response to November 2023 Illegal Harms Consultation, p.20.
[1872] Global Network Initiative response to November 2023 Illegal Harms Consultation, p.4; Safe Space One response to November 2023 Illegal Harms Consultation, p.20.
[1873] British and Irish Law, Education, and Technology Association (BILETA) response to November 2023 Illegal Harms Consultation, p.20; Mencap response to November 2023 Illegal Harms Consultation, pp.16-17; Nexus response to November 2023 Consultation, pp.21-22; OneID response to November 2023 Consultation, p.6; Sanders, T. response to November 2023 Illegal Harms Consultation, p.17.
[1874] Cifas response to November 2023 Illegal Harms Consultation, p.20; [✂]; National Trading Standards eCrime Team response to November 2023 Illegal Harms Consultation, p.16; South East Fermanagh Foundation (SEFF) response to November 2023 Illegal Harms Consultation, p. 21.
[1875] Clean Up the Internet response to November 2023 Illegal Harms Consultation, p.6; National Society for the Protection of Children (NSPCC) response to November 2023 Illegal Harms Consultation, p.46.
[1876] Airbnb response to November 2023 Illegal Harms Consultation, p.12; [✂]; Meta and Whatsapp response to November 2023 Illegal Harms Consultation, p.18; Microsoft response to November 2023 Consultation, p.8; Snap response to November 2023 Illegal Harms Consultation, p.5. We note that Snap also made a similar point in response to the May 2024 Consultation on Protecting Children from Harms Online, p.13; Spotify response to November 2023 Consultation, p.12; techUK response to November 2023 Illegal Harms Consultation, p.18; Ukie response to November 2023 Illegal Harms Consultation, p.32; UK Safer Internet Centre (UKSIC) response to November 2023 Illegal Harms Consultation, p.34.
[1877] Cifas response to November 2023 Consultation, p.6; Wikimedia Foundation response to November 2023 Illegal Harms Consultation, p.16.
[1878] Google response to November 2023 Illegal Harms Consultation, p.69; Meta and WhatsApp response to November 2023 Consultation, p.34.

# We consider the overall package proportionate

## Each individual measure delivers distinct benefits

13.5 We recommend some measures aimed at tackling all illegal harms rather than targeting specific harms. They aim to establish robust governance processes, content (or search) moderation systems, clear and accessible reporting and complaints and terms of service functions. As a package, these measures act in a complementary way to help keep users safe and removing any single measure would diminish the overall effectiveness of the package:

- Individual measures work in different ways to reduce the risk of illegal harms. For example, governance and moderation measures largely focus on ensuring the effective operation of various internal systems and processes that contribute to user safety, while some other measures related to reporting and terms of service address issues related to end-user functionalities and experience, such as ease of use and accessibility.
- Some measures are complementary such that their combined benefit is higher than when considering measures individually. For example, better governance can support effective implementation of safety measures in general, and better reporting and complaints functions can lead to more effective content moderation.

13.6 We also recommend a measure that applies to multi-risk services with content recommender systems that carry out on-platform testing. It should help them reduce the amount of specific types of illegal content disseminated by recommender algorithms. Therefore, this measure provides distinct benefits as it addresses a specific risk factor that other measures would not sufficiently address.

13.7 Other measures target specific harms, including measures for child sexual exploitation and abuse (CSEA), terror, harassment and fraud. We consider the benefits of these measures do not substantially overlap. First, some measures address different kinds of illegal harms. Second, even where more than one measure targets the same harm, they focus on different aspects of it and deliver distinct benefits. For example, one of our anti-grooming measures makes it harder for perpetrators to find and contact children, and another provides children with the information they need to make informed decisions to better manage their risks. Moreover, the magnitude of the challenge is large and even the package of measures will not deal comprehensively with illegal content, especially for this first set of codes on which we intend to build in the future.

13.8 Overall, we conclude that there is not significant duplication between the measures we have recommended and in many cases they are complementary. Therefore, each measure has distinct benefits and contributes to reducing the risks of illegal harms (even in addition to the other measures in the package).

## We recommend very few measures for smaller low-risk services, beyond the specific requirements in the Act

13.9 We acknowledge the concerns raised by some respondents that our proposed measures could have a detrimental effect on providers of smaller low-risk services because of the regulatory burden. This could include services run by small and micro businesses, as well as those run by individuals and charities on a non-commercial basis. It would almost certainly be against users' interests if the regulatory burden on such providers leads to degradation in service quality or user experience, or to them ceasing to operate in the UK. This is because users would lose access to services they currently use (or experience a lower quality of service) without necessarily receiving any offsetting benefit from a reduction in their risk of harm due to the low-risk nature of the services. As one stakeholder noted, **"**The internet has revolutionised communication; it would be very sad if the harmless parts of it were destroyed by measures to control the harmful parts."[1879]

13.10 Therefore, we have reduced the number of measures that apply to smaller low-risk services compared to our consultation proposals. Most of the measures imposed on smaller low-risk services are a direct result of specific requirements in the Act, over which we have limited discretion. We have imposed very few measures beyond that.[1880] The combined costs of these additional measures over which we have discretion is expected to be very low.

## The regulatory burden for other services is proportionate

13.11 We recommend more demanding measures for providers of **smaller risky services**, compared to those that are low-risk. We recognise that the combined impact of the cost of these measures could be very significant for some services. Some small and micro businesses may struggle to resource the recommendations. It is even possible that some services may cease to operate in the UK. Even if this were to happen, we do not consider it would mean that the measures are disproportionate, given the risks present on their services. Given we assess each measure as proportionate and the benefits of the measures do not overlap to a significant extent, we consider the combined impact of the measures to be proportionate.

13.12 We recommend even more demanding measures for providers of some **large services**.[1881] The overall package of measures could entail significant costs. However, we generally expect providers of large services to have the resources to undertake these measures. Even if some providers of large risky services have more limited resources, the package of measures would still be proportionate, for the same reasons as for providers of smaller risky services. Where large services are low-risk, most of the measures recommended for such services have sufficient flexibility to allow the provider to implement without high cost.

## Conclusion

13.13 Based on the above, we consider the overall package of recommended measures to be proportionate. Our assessment shows that each measure has distinct benefits and

[1879] Bolton, C. response to November 2023 Consultation, p. 11.

[1880] The main additional measure we impose over which we have more discretion is for providers to name a person accountable for compliance with illegal content safety duties. This applies to providers of all services. Providers of smaller, low-risk U2U services must also remove any accounts operated by proscribed terror organisations. If there were a general search service that was both small and low-risk, it would be required to take steps to remove URLs identified as hosting child sexual abuse material ('CSAM') from search results. There is also a measure that sets out the conditions under which any U2U or search service can disregard manifestly unfounded complaints. This only applies where providers choose to do this, and we envisage it being relevant for services that receive a large volume of complaints rather than small services.

[1881] We explain in 'Our approach to developing Codes measures' why we decide to maintain our consultation proposal on the definition of large services and our general approach to apply more measures to large services compared to smaller ones.

contributes to reducing the risks of illegal harms, even in addition to the other measures in the package. While the overall cost could be very significant for some services, we consider it proportionate given the risks of users encountering illegal harms on the services and the incremental benefit of each measure in the package in reducing these risks. Therefore, we consider that the package of measures is justified and proportionate, consistent with the assessments we have set out for each measure in earlier chapters.

# 14. Statutory Tests

## What is this chapter about?

In designing our Codes, the Online Safety Act requires us to have regard to a number of principles and objectives, set out in Schedule 4 to the Act. The Communications Act 2003 also places a number of duties on us in carrying out our functions.

In this chapter we set out the matters to which we must have regard under the Online Safety Act and the Communications Act, and explain the reasons why we think the recommendations in our illegal content Codes of Practice meet them. We provide further information regarding Ofcom's duties relating to the preparation of our Codes in our introduction to the Statement, our Legal Framework (Annex 2), and Annex 4, in which we set out our Equality Impact Assessment and Welsh language assessments.

## Background

14.1 In designing our Codes, the Online Safety Act requires us to have regard to a number of principles and objectives, set out in Schedule 4 to the Act. The Communications Act 2003 also places a number of duties on us in carrying out our functions, including requiring us to have regard to the risk of harm to citizens presented by content on regulated services.

14.2 In Chapters 2 to 12 in this volume, we set out our recommended Codes measures; an overview of these recommendations can be found in 'Overview of Illegal Harms', and our 'Combined Impact Assessment' of the measures can be found in Chapter 13 of this volume. The measures themselves can be found in full in the 'Illegal Content Codes of Practice for U2U services' and 'Illegal Content Codes of Practice for search services'. We provide further information regarding Ofcom's duties relating to the preparation of our Codes in our 'Legal Framework Overview' in Annex 2.

14.3 We consider that our recommendations meet the requirements set out in Schedule 4 to the Online Safety Act and section 3 of the Communications Act. In this chapter, we take each of the requirements in turn and set out how we have met them.

## Summary of stakeholder feedback[1882]

14.4 Stakeholders, including financial organisations, civil society, and other large service providers, expressed broad support for Ofcom's recommendations for the Codes in the November 2023 Illegal Harms Consultation.[1883]

[1882] Note that responses listed in this summary are not exhaustive, and further responses can be found in Annex 1.

[1883] Are, C response to November 2023 Illegal Harms Consultation, p.21; Local Government Association response to November 2023 Illegal Harms Consultation, p.17; Match Group response to November 2023 Illegal Harms Consultation, p.20-21; Nexus response to November 2023 Illegal Harms Consultation, p.22; Protection Group International response to November 2023 Illegal Harms Consultation, p.4; Segregated Payments Ltd response to November 2023 Illegal Harms Consultation, p.16.

14.5 Stakeholders responded to us in detail on the proportionality of our measures and their impact on human rights. We address these concerns in ‘Our approach to developing Codes measures’, and the chapters on our individual measures and do not repeat them here.

14.6 Some stakeholders raised concerns about the **technical feasibility** of specific measures. We address these points in the specific chapters. (See our Content Moderation and Search Moderation chapters).[1884]

14.7 The Molly Rose Foundation said the disconnect between evidence of harm in risk profiles and register of risks and mitigation for those risks in the Codes represents a failure by Ofcom to meet safety objectives specified in schedule 4 (3).[1885] We address this comment in ‘Our approach to developing Codes measures’.

14.8 The Independent Reviewer of Terrorism Legislation argued that we had not complied with our duty to ensure that the Code of Practice for terrorism content provides a higher standard of protection for children than for adults.[1886] We address this under the following sub-headings of this chapter: user to user services, 4(a)(vi); and search services 5(a)(v).

## Appropriateness and principles

14.9 As required by section 3 of the Communications Act 2003, 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.

14.10 As required by paragraph 1 of Schedule 4 to the Online Safety Act, Ofcom has considered the appropriateness of provisions of 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 applying some Codes recommendations to services of different kinds, sizes and capacities.[1887]

14.11 Ofcom has had regard to the following principles in Schedule 4, as follows:

**Paragraph 2(a)**: providers of Part 3 services must be able to understand which provisions of the code of practice apply in relation to a particular service they provide.[1888]

a) Ofcom has clearly identified in our Codes which measures apply to what types and sizes of services, for the reasons given in each relevant chapter of this statement.

**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.[1889]

[1884] The Internet Society (at p.11 of its response) said that Ofcom’s identification of encrypted messaging as a risk factor was inconsistent with the Act, because the safety duty does not apply to private communications. We do not think this is a correct interpretation of the Act. Proactive technologies cannot be recommended in Codes for content communicated privately, but other measures can. The Internet Society response to November 2023 Illegal Harms Consultation, p.11

[1885] Molly Rose Foundation response to November 2023 Illegal Harms Consultation, p.29.

[1886] Independent Reviewer of Terrorism Legislation response to November 2023 Illegal Harms Consultation, pp.4-5.

[1887] See also section 3(4A)(d) Communications Act 2003.

[1888] See also section 3(4A)(c) Communications Act 2003.

[1889] See also section 3(4A)(c) Communications Act 2003.

a) Having regard to the need for it to be clear to providers of regulated services how they may comply with their duties, Ofcom has aimed to be as clear and detailed as possible in our Codes, consistent with acting proportionately.

b) Some stakeholders said that our Codes required further clarification and detail.[1890] We have sought to be sufficiently detailed and precise, and made changes to our Codes to ensure they are sufficiently clear. In particular:

   i) We have amended some of our measures on governance so as to be clearer that we expect management oversight of the risks which remain after our measures are adopted.

   ii) We have amended our measure on prioritisation of appeals to be clearer that it does not require all providers to have regard to every factor in every decision.

   iii) We have amended our measure on recommender systems so as to be clear that it does not apply to product recommender systems.

   iv) We have amended our Codes to be clearer about how they apply to UK users.

14.12 Some stakeholders asked for more explanation of our use of the word "swiftly" in Measure ICU C2. We have decided to keep the same definition. See section 'Swiftly taking down content of which providers are "aware"' in chapter 2 of this volume: 'Content moderation' where we address these stakeholder responses.[1891]

14.13 techUK recommended that we provide enough time when finalising our Codes to allow providers time to clarify if their services are in scope and consult with providers in complex supply chains.[1892] We consider that there has been enough time since the Act was passed for providers to determine whether they are in scope and to prepare to meet their obligations under the safety duty.

> **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;

a) Ofcom is recommending measures many of which we know to be in widespread use in the sector. Ofcom has clearly identified in our Codes which measures apply to what types and sizes of services, for the reasons given in each relevant chapter of this statement.

b) We have considered proportionality and technical feasibility, where appropriate, as part of our impact assessment. We do so in our consultation and for this statement. This includes taking into account evidence of current practice by user-to-user and search service providers which are already taking steps that are similar or related to measures that we propose. We consider effectiveness, costs, rights impacts, and other relevant factors in our assessment of proportionality. The more demanding measures, we recommend for services that pose greater risk of harm, even if they are smaller services. At the same time, certain measures are recommended for large services only, based on

[1890] British and Irish Law, Education and Technology Association (BILETA) said some definitions were not "concrete" enough and may lead to inconsistent enforcement or legal disputes. British and Irish Law, Education and Technology Association (BILETA) response to November 2023 Illegal Harms Consultation, p.20

[1891] BILETA response to November 2023 Consultation, p.20.

[1892] techUK response to November 2023 Illegal Harms Consultation 2023, p.27.

proportionality considerations including with respect to the capacity of smaller services to implement them.

c) Some stakeholders did not agree with applying measures to 'large' services or with our definition of 'large services'. Some stakeholders also disagreed with our position that some measures should not apply to smaller services.[1893] We have not changed our position on these points. See section 'Large Services' in 'Our approach to developing Codes measures' where we address these stakeholder responses.

d) Some stakeholders thought our measures would lead to over-compliance and over-moderation on some services as well as over-targeting users.[1894] See paragraph 1.69 in 'Our approach to developing Codes measures' where we address these stakeholder responses. See also the 'Freedom of expression and risk of over-takedown' section in Volume 3 in the ICJG.

e) Stakeholders thought some of our measures were not technically feasible for all platforms.[1895] See our Content Moderation and Search Moderation chapters, where we address these points.

**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.

a) Ofcom has identified in our reasoning the harms which our recommendations would address, and explained why we consider each measure is proportionate in the light of those harms. As required by section 3(4A)(b)(ii) of the Communications Act 2003, in considering proportionality we have had regard to the severity of the potential harm as well as the level of risk of harm, as identified in our Register of Risks. Where appropriate, Ofcom has clearly identified in our Codes which measures would apply to what types and sizes of services, for the reasons given in each relevant chapter of this statement.

b) Having had regard to the desirability of promoting the use by providers of regulated services of technologies which are designed to reduce the risk of harm to citizens presented by content on regulated services,[1896] to the desirability of encouraging investment and innovation in this market, and to the seriousness of the harms concerned, Ofcom has, in particular, recommended the use of certain kinds of technologies where proportionate to the risk of harm from CSAM.

c) We recognise that technologies used by services which are designed to reduce the risk of harm to citizens are constantly evolving due to technological advancements that improve their efficiency and effectiveness. While our approach recommends the use of automated content moderation systems as a general type of technology, it does not make recommendations regarding the use of specific technologies or use of specific inputs (such as hash databases or URL lists provided by a specified third party). See the Automated content moderation chapter in this volume where we address the use of these technologies.

[1893] Microsoft response to November 2023 Illegal Harms Consultation, pp.8, 9 and 19.

[1894] Are, C response to November 2023 Consultation, p.21.

[1895] BILETA response to November 2023 Consultation, p.20; The Internet Society response to November 2023 Consultation, pp.12-13.

[1896] Section 3(4A)(a), (d) and (e) Communications Act 2003.

# Ofcom's general duties and the online safety objectives

## U2U services

14.14 As required by paragraph 3 of Schedule 4 to the Online Safety Act, Ofcom has also ensured that the recommendations are compatible with the pursuit of the applicable online safety objectives for U2U services as follows:

> **Paragraph 4(a)(i)**: a service should be designed and operated in such a way that the systems and processes for regulatory compliance and risk management are effective and proportionate to the kind and size of service.

a) In Volume 1: chapter 5: 'Governance and accountability', Ofcom has set out the governance measures which we recommend having regard, amongst other things, to the kind and size of service. We consider these to be compatible with this objective.

> **Paragraph 4(a)(ii)**: a service should be designed and operated in such a way that the systems and processes are appropriate to deal with the number of users of the service and its user base.

a) As set out in our overview, we have considered the size of services in our assessment of whether the recommendation of certain measures is proportionate; in Volume 1: chapter 5: 'Governance and accountability', and the following chapters in this volume: chapter 2: 'Content moderation', chapter 4: 'Automated content moderation', chapter 6: 'Reporting and complaints', and chapter 12: 'User controls'*,* Ofcom has set out the systems and processes measures which we recommend having regard, amongst other things, to the number of users of the service and its user base. We consider these to be compatible with this objective.

> **Paragraph 4(a)(iii)**: a service should be designed and operated in such a way that United Kingdom users (including children) are made aware of, and can understand, the terms of service.

a) In chapter 10 of this volume: 'Terms of service/Publicly available statements (ToS/PAS)' we set out our reasoning for a recommendation which we consider compatible with this objective. In making this recommendation, we have also considered our duty to have regard to the extent to which providers of regulated services demonstrate, in a way that is transparent and accountable, that they are complying with their duties set out in the Act.[1897]

> **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.

a) In chapter 6: 'Reporting and Complaints', and chapter 8: 'U2U Settings, Functionalities, and User Support' of this volume we have included recommendations which we consider are compatible with this objective.

[1897] Section 3(4A)(f) Communications Act 2003.

**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.[1898]

a) Our Codes of Practice for all harms provide more protection to children than to adults.
b) In particular, in chapter 8 of this volume: 'U2U Settings, Functionalities, and User Support' we have included recommendations which we consider would be compatible with this objective.
c) Providers of services likely to be accessed by children must (i) provide extra information through a reporting function/tool easily accessible prior to the submission of a complaint Measure ICU D2 and D3, and (ii) send further information about how the complaint will be handled Measure ICU D5.
d) We have also clarified in our measure on prioritisation policies for content moderation that whether a harm affects a child is an aspect of the severity of harm.
e) Our complaints and reporting and content moderation measures are part of each of our Illegal Content Codes, including our terrorism Code. However, more generally we consider it appropriate and more effective to deal with the promotion of violent and hateful content in our Children's Safety Codes of Practice, as this does not require providers to apply complex and potentially contentious UK legal definitions of 'terrorism'. We are considering whether we need to further explain the risks of radicalisation in our Childrens Harms Guidance.
f) As set out in the 'Iterative approach' section of 'Our approach to developing Codes measures', our strategy is to implement our first Codes as soon as possible and then build on them over time.

**Paragraph 4(a)(vii)**: a service should be designed and operated in such a way that the different needs of children at different ages are taken into account.

a) In chapter 10: 'Terms of Service', and chapter 8: 'U2U Settings, Functionalities, and User Support' of this volume we set out how we have had regard to the different needs of children at different ages. We have included recommendations which we consider would be compatible with this objective.

**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.

a) In chapter 11 of this volume: 'User Access' we set out why we do not consider it appropriate to restrict access to services generally by adults. We explain the measure we have recommended to limit the activities of proscribed organisations. In chapter 12 of this volume: 'User Controls' we set out the steps we expect a service to take if it purports to offer a verification scheme for users.

**Paragraph 4(a)(ix)**: a service should be designed and operated in such a way that there are adequate controls over access to, and use of, the service by children, taking into account use of the service by, and impact on, children in different age groups.

a) In Chapter 8 (U2U Settings, Functionalities, and User Support) we have explained our recommendations which we consider would be compatible with this objective, and

[1898] See also section 3(4A)(b) Communications Act 2003.

explained how we have taken into account use of the service by, and impact on, children in different age groups.

**Paragraph 4(b)**: a service should be designed and operated so as to protect individuals in the United Kingdom who are users of the service from harm, including with regard to—

- algorithms used by the service,
- functionalities of the service, and
- other features relating to the operation of the service.

a) All our recommendations seek to protect users from harm. In particular, in Volume 1: chapter 5: 'Governance and Accountability', and the following chapters of this volume: chapter 2: 'Content Moderation', chapter 4: 'Automated Content Moderation', chapter 6: 'Reporting and Complaints', chapter 8: 'U2U Settings, Functionalities, and User Support', and chapter 7: 'Recommender Systems', we have included recommendations which we consider would be compatible with this objective.

14.15 We are not at this stage recommending measures relating to paragraph 4(a)(v) given it is specific to Category 1 services only.

## Search services

14.16 As required by paragraph 3 of Schedule 4 to the Online Safety Act, Ofcom has ensured that the recommendations are compatible with the pursuit of the applicable online safety objectives for search services as follows:

**Paragraph 5(a)(i)**: a service should be designed and operated in such a way that the systems and processes for regulatory compliance and risk management are effective and proportionate to the kind and size of service.

a) In Volume 1: chapter 5: 'Governance and Accountability', Ofcom has set out the governance measures we have decided to recommend having regard, amongst other things, to the kind and size of service. We consider these to be compatible with this objective.

**Paragraph 5(a)(ii)**: a service should be designed and operated in such a way that the systems and processes are appropriate to deal with the number of users of the service and its user base.

a) In Volume 1: chapter 5: 'Governance and Accountability', and the following chapters of this volume: chapter 3: 'Search Moderation', chapter 5 'Automated Search Moderation', and chapter 9: 'Search Design, Functionalities, and User Support', Ofcom has set out the systems and processes measures which we recommend having regard, amongst other things, to the number of users of the service and its user base. We consider these to be compatible with this objective.

**Paragraph 5(a)(iii)**: a service should be designed and operated in such a way that United Kingdom users (including children) are made aware of, and can understand, the publicly available statement referred to in sections 27 and 29.

a) In chapter 10 of this volume: 'PAS' we have included a recommendation which we consider is compatible with this objective. In making this recommendation, we have also considered our duty to have regard to the extent to which providers of regulated services demonstrate, in a way that is transparent and accountable, that they are complying with their duties set out in the Act.

**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.

a) In the following chapters of this volume: chapter 3: 'Search Moderation', chapter 5: 'Automated Search Moderation', and chapter 9: 'Search Design, Functionalities, and User Support' we have included recommendations which we consider are compatible with this objective. For example, we have included a search design measure for CSAM content warnings so that they are suitable and comprehensible in content and tone for as many users as possible, including children Measure ICS F2.

**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.[1899]

a) Having had careful regard to the need for a higher level of protection for children than for adults, in chapter 5 of this volume: 'Automated Search Moderation' we have included a recommendation which we consider is compatible with this objective.
b) Providers of services likely to be accessed by children must (i) provide extra information through a reporting function/tool easily accessible prior to the submission of a complaint Measures ICS D2, and (ii) send further information about how the complaint will be handled Measure ICS D4.
c) We have also clarified in our measure on prioritisation policies for search moderation that whether a harm affects a child is an aspect of the severity of harm.
d) Our complaints and reporting and content moderation measures are part of each of our Illegal Content Codes, including our terrorism Code. However, more generally we consider it appropriate and more effective to deal with the promotion of violent and hateful content in our Children's Safety Codes of Practice, as this does not require providers to apply complex and potentially contentious UK legal definitions of 'terrorism'. We are considering whether we need to further explain the risks of radicalisation in our Childrens Harms Guidance.
e) As set out in the 'Iterative Approach section' of 'Our approach to developing Codes measures', our strategy is to implement our first Codes as soon as possible and then build on them over time.

**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.

a) In chapter 6 of this volume: 'Reporting and Complaints', and chapter 10 of this volume: 'PAS' we set out how we have had regard to the different needs of children at different ages.

**Paragraph 5(b)**: a service should be assessed to understand its use by, and impact on, children in different age groups.

[1899] See also section 3(4A)(b) Communications Act 2003.

a) We have had regard to the needs of children of all ages, but consider that this objective is better advanced via the risk assessment duties. We consider our recommendations in relation to illegal content, in particular those relating to governance, are compatible with it.

**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

a) In Volume 1: chapter 5: 'Governance and Accountability', and the following chapters in this volume: chapter 3: 'Search Moderation', chapter 5: 'Automated Search Moderation', and chapter 9: 'Search Design, Functionalities, and User Support' we have included recommendations which we consider would be compatible with this objective.

# Content of Codes of Practice

## U2U services

14.17 Codes of practice that describe measures recommended for the purpose of compliance with a duty set out in section 10(2) or (3) (illegal content) must include measures in each of the areas of a service listed in section 10(4). This provision applies to the extent that inclusion of the measures in question is consistent with:

a) Ofcom's duty to consider the appropriateness of provisions of the Codes of practice to different kinds and sizes of Part 3 services and to providers of differing sizes and capacities;

b) the principle that the measures described in the Codes of practice must be proportionate and technically feasible: measures that are proportionate or technically feasible for providers of a certain size or capacity, or for services of a certain kind or size, may not be proportionate or technically feasible for providers of a different size or capacity or for services of a different kind or size; and

c) the principle that the measures described in the Codes 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.

14.18 Ofcom has made recommendations for U2U services in each of the areas of a service listed in section 10(4) as follows:

a) regulatory compliance and risk management arrangements – see Volume 1: chapter 5: 'Governance and Accountability',

b) design of functionalities, algorithms and other features – see chapter 8 of this volume: 'U2U Settings, Functionalities, and User Support', and chapter 7 of this volume 'Recommender Systems',

c) policies on terms of use – see chapter 10 of this volume: 'ToS', and chapter 2 of this volume: 'Content Moderation',

d) policies on user access to the service or to particular content present on the service, including blocking users from accessing the service or particular content – see chapter 11 of this volume: 'User Access',
e) content moderation, including taking down content – see chapter 2 of this volume: 'Content Moderation', and chapter 4 of this volume: 'Automated Content Moderation',
f) functionalities allowing users to control the content they encounter – see chapter 8 of this volume: 'U2U Settings, Functionalities, and User Support', chapter 12 of this volume: 'User Controls',
g) user support measures – see chapter 8 of this volume: 'U2U Settings, Functionalities, and User Support',
h) staff policies and practices – see Volume 1: chapter 5: 'Governance and Accountability', and chapter 2 of this volume: 'Content Moderation'.

14.19 We designed the measures in line with paragraph 10(1)-(3) of Schedule 4 of the Act which requires measures described in a Codes of practice which are recommended for the purpose of compliance with any of the relevant duties, to be designed in the light of the following principles:

a) The importance of protecting the rights of users to freedom of expression within the law.
b) The importance of protecting the privacy of users.

14.20 We set out further information about our consideration of human rights below.

14.21 All the measures we recommend in the Codes relate only to the design or operation of a Part 3 service (a) in the United Kingdom, or (b) as it affects United Kingdom users of the service.

## Search services

14.22 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) must include measures in each of the areas of a service listed in section 27(4). This provision applies to the extent that inclusion of the measures in question is consistent with:

a) Ofcom's duty to consider the appropriateness of provisions of the Codes of practice to different kinds and sizes of Part 3 services and to providers of differing sizes and capacities;
b) the principle that the measures described in the Codes of practice must be proportionate and technically feasible; and
c) the principle that the measures described in the Codes 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.

14.23 Ofcom has made recommendations for search services in the following areas of a service listed in section 27(4) as follows:

a) regulatory compliance and risk management arrangements – see Volume 1: chapter 5: 'Governance and Accountability',
b) design of functionalities, algorithms and other features relating to the search engine – see the following chapters in this volume: chapter 6: 'Reporting and Complaints',

chapter 5: ‘Automated Search Moderation’, and chapter 9: ‘Search Design, Functionalities, and User Support’,

c) user support measures – see chapter 9: ‘Search Design, Functionalities, and User Support’,

d) staff policies and practices – see Volume 1: chapter 5: ‘Governance and Accountability’, and chapter 3 of this volume: ‘Search Moderation’.

14.24 For the reasons set out in the relevant sections, Ofcom did not consider it appropriate or proportionate at this stage to make recommendations for search services in relation to one area of a service listed in section 27(4): functionalities allowing users to control the content they encounter in search results. We consider risks relating to these areas will be better addressed through our work on protection of children.

14.25 We have designed the measures in line with paragraph 10(1)-(3) of Schedule 4 of the Act, in the light of the following principles:

a) The importance of protecting the rights of users and (in the case of search services or combined services) interested persons to freedom of expression within the law.

b) The importance of protecting the privacy of users.

14.26 All the measures we recommend in the Codes 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.

## Human rights

14.27 As set out in Chapter 1 of this volume: ‘Introduction to the volume’ and the ‘Purpose of Codes of Practice section’ in ‘Our approach to developing Codes measures’ Ofcom has had careful regard to the right to freedom of expression and the right to respect for private and family life in making these recommendations.

14.28 Decisions at both a domestic level and before the European Court of Human Rights make clear the scope for restrictions on freedom of expression is likely to be especially limited in two overlapping fields, namely political speech and on matters of public interest. Accordingly, a high level of protection of freedom of expression will normally be accorded to these types of speech, with the authorities having a particularly narrow margin of appreciation. Intellectual and educational speech and artistic speech and expression are also considered deserving of protection under Article 10, while "mere abuse" (i.e. gratuitously offensive speech that does not contribute to public debate) attracts the lowest level of protection. Hate speech is afforded no protection under Article 10. The measures we are recommending are likely to affect all these types of expression.

14.29 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.

14.30 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).

14.31 The need for any restriction of these rights must be construed strictly and established convincingly. Any interference must be prescribed by law; pursue a legitimate aim (as set out in Articles 8(2), 9(2), 10(2) and 11(2)); and be necessary in a democratic society - in other words, it must be proportionate to the legitimate aim pursued and corresponding to a pressing social need.

14.32 In passing the Act, Parliament has set out in legislation the interferences prescribed by law and which it has judged to be necessary in our democratic society. These relate to the protection of users from harm they may experience on regulated services, particularly from exposure to illegal content, or where user to user services are used for the facilitation or commission of priority offences. The relevant legitimate aims that Ofcom may act in pursuit of in the context of our functions under the Act therefore include the prevention of crime and disorder, public safety and the protection of health or morals, and the protection of the rights and freedoms of others.

14.33 Where we have identified the potential for interference with ECHR rights, we have carried out a careful analysis of where the interference is proportionate. Our starting point has been to recognise that Parliament has determined that regulated providers must take proportionate measures to fulfil their duties to protect users from illegal content and, where relevant, to address the risk of services being used for the commission or facilitation of priority offences. Such measures will necessarily have an impact on the experiences of those who are using these services, in particular by significantly limiting or preventing users' exposure to such content. Any errors in identifying content that is illegal, and any steps taken to limit use of a service by suspected criminals or the visibility of their content, could impact their rights to freedom of expression, and in some cases, their rights to freedom of religion or belief and freedom of association.

14.34 They will also, to some extent, have impacts on users' rights to privacy and their data protection rights, insofar as they would require their private information or personal data to be processed for the measures to work properly.

14.35 To the extent that such interferences can be seen as a direct result of the duties imposed on services, and Ofcom, by Parliament, and are required to achieve the legitimate objective of securing adequate protections for users from harm, we consider that a substantial public interest exists in these outcomes.

14.36 However, in line with our obligations under the Human Rights Act, we also seek to secure that any such interference with users' rights to freedom of expression and privacy, or other relevant rights, is proportionate to the legitimate objectives pursued. Where appropriate we have explained why the relevant restriction is justified, and have sought to build into our Codes measures appropriate safeguards to protect those rights.

14.37 In doing so, among other things, we have carefully considered whether other, less intrusive measures are available that might adequately mitigate the harms faced by users on regulated services.

14.38 Overall, we have sought to strike a fair balance between securing adequate protections for users from harm (and their rights in respect of this) and the ECHR rights of users, other interested persons (including for example, persons who host websites and 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 services, as relevant.[1900] In other words, we are concerned to ensure that the degree of interference with ECHR rights is outweighed by the benefits secured in terms of protecting users from harm.

14.39 In seeking to achieve this fair balance, we consider that the Act and the protection it gives to individuals against harms of various kinds reflect the decision of the UK Parliament that UK users should be proportionately protected from all the harms concerned. In doing so, Parliament has enshrined in UK law the rights of UK users - including their human rights - to be protected from those harms. In weighing up whether the measures we are proposing are proportionate, we 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 Register of Risks.

14.40 We considered each measure separately in the preceding chapters. We do not think the analysis is different taking all the measures collectively. Overall, we are satisfied that our recommendations are compatible with human rights.

[1900] This reflects the fourth limb of the '*Bank Mellat* test'. In response to our consultation, OSA Network (Annex C of its response) argued that we should have focused more on the rights of those harmed by regulated online content. We consider that the weighing up exercise described here includes those rights.