

# Protecting people from illegal harms online

---

Volume 4:  
How to mitigate the risk of illegal harms –  
the illegal content Codes of Practice

## Consultation

Published: 9 November 2023

Closing date for responses: 23 February 2024



# Contents

---

## Section

11. Introduction: Our approach to the illegal content Codes of Practice .....	3
12. U2U content moderation.....	19
13. Search moderation .....	59
14. Automated Content Moderation (U2U).....	89
15. Automated Search Moderation .....	153
16. Reporting and complaints.....	171
17. Terms of service and publicly available statements.....	216
18. U2U default settings and support for child users .....	229
19. Recommender system testing (U2U) .....	265
20. Enhanced User Control (U2U).....	280
21. User access .....	312
22. Search features, functionalities and user support .....	336
23. Cumulative Assessment of Proposed Measures .....	356
24. Statutory Tests.....	364

# 11. Introduction: Our approach to the illegal content Codes of Practice

## What is this chapter about?

This volume focuses on the steps we propose to recommend services should take to mitigate the risk of illegal harms. These recommendations are captured in our illegal content Codes of Practice ('Codes'). This chapter describes the approach we propose to take to developing the Codes. Subsequent chapters in this volume describe the specific measures we are proposing to include in the Codes.<sup>1</sup>

Our proposed recommendations for Codes cover core areas encompassing all areas of the design, operation and use of an in-scope service. We propose separate Codes for each of U2U and search services, with some measures being common to both.<sup>2</sup>

We believe that these first Codes represent a strong basis on which to build a more comprehensive suite of recommended measures to reduce the risk of harm to users over the longer term. In this vein, our first Codes aim to capture existing good practice within industry and set clear expectations on raising standards of user protection, especially for services whose existing systems are patchy or inadequate. Each proposed measure has been impact assessed, considering harm reduction, effectiveness, cost and the impact on rights.

We have carefully considered the proportionality and the cumulative impact of our proposals. Given the range and diversity of services in scope, we are not taking a 'one size fits all' approach. We propose a small number of measures for all U2U services, and a similar set of measures for all search services. Beyond these, many of the other measures depend on the risks the service has found in its latest illegal content risk assessment and the size of the service.

Some measures are targeted at addressing the risk of certain kinds of offences, such as CSAM, grooming and fraud. Other measures are intended to address a wide range of offences. We intend these measures to apply to services that face significant risks for offences in general.

Services that decide to implement measures recommended to them for the kinds of illegal harms and their size or level of risk indicated in our Codes of Practice will be treated as complying with the relevant duty. This means that Ofcom will not take enforcement action against them for breach of that duty.

## What are we proposing?

We provide a full list of the measures we propose in Codes and a breakdown of which types of services we would expect to do them in the 'tear sheet' document accompanying the consultation. In this chapter we make a number of overarching proposals regarding our approach:

---

<sup>1</sup> This is with the exception of measures relating to governance and accountability, which are contained in Chapter 8 of volume 3, as volume 3 relates to how services should assess the risks of online harm. The proposal in Chapter 8 does however form part of the Codes we propose.

<sup>2</sup> We have decided to present each of the U2U and search Codes of Practice as a single document to aid readability and reduce duplication and cross-referencing between the various Codes we are obliged to produce under the Act, namely Codes of practice for Schedule 5 (terrorism) offences, Schedule 6 (child sexual abuse and exploitation offences), and the other offences.

- Some of the measures we are proposing target specific kinds of illegal harms. We propose to apply the most onerous harm-specific measures in our Codes only to services which are large and/or medium or high risk for the specific kinds of illegal harm we are targeting.
- Some of the measures we are proposing target a wide range of online harms. We propose to apply the most onerous of these measures in our Codes only to services which are large and/or multi-risk.
- We propose to define a service as large where it has an average user base greater than 7 million per month in the UK, approximately equivalent to 10% of the UK population.
- We propose to define a service as multi-risk where it is high or medium risk for at least two kinds of illegal harms.

### Why are we proposing this?

Focusing the most onerous measures on services which are large and/or medium or high risk will help ensure that the impact of the regulations is proportionate. All else being equal, the benefits of a measure will be greater when they are applied to services with a bigger user base. At the same time, all else being equal, the benefits of a medium or high-risk service implementing a measure will generally be higher than the benefits of a low-risk service implementing a measure.

As we explain in more detail below, where services pose a high risk of causing harm, we apply more onerous measures to them even when they are small. Whilst there is sometimes a correlation between size and risk, in the case of some harms (for example grooming) small services can pose a high risk of harm. Where risks are very high, it is important that people are afforded protection even when the services they are using are relatively small.

We consider larger services will tend to be better able to bear the costs of the more onerous measures than smaller services. Our definition of large closely mirrors the definition of large services taken by the EU in the Digital Services Act. We consider it important to broadly align our approach to determining larger services with other international regimes where possible, to reduce the potential burden of regulatory compliance for services.

### What input do we want from stakeholders?

- Do you have any comments on our overarching approach to developing our illegal content Codes of Practice?
- Do you agree that in general we should apply the most onerous measures in our Codes only to services which are large and/or medium or high risk?
- Do you agree with our definition of large services?
- Do you agree with our definition of multi-risk services?
- Do you have any comments on the draft Codes of Practice themselves?<sup>3</sup>
- Do you have any comments on the costs assumptions set out in Annex 14, which we used for calculating the costs of various measures?

---

<sup>3</sup> Please see Annexes 7 and 8 to find our draft Codes of Practice.

## Background to the Illegal Content Codes of Practice

---

- 11.1 This chapter contains:
- The background to the Illegal Content Codes of Practice;
  - Our key proposed recommendations;
  - Our approach to developing recommended measures;
  - The structure of this volume.
- 11.2 When finalised, Ofcom’s Codes of Practice (**‘Codes’**) will set out the measures we will recommend for services to comply with their safety duties. This volume of the consultation lays out our proposed approach to designing and presenting our Illegal Content Codes of Practice and the measures we propose to recommend. We have published the draft Illegal Content Codes of Practice in Annexes 7 and 8.
- 11.3 Under the Online Safety Act, Ofcom is required to prepare and issue three sets of Codes for Part 3 services (U2U and search services), namely a Code covering terrorism content, relating to the offences set out in Schedule 5, a Code on CSEA content relating to the offences set out in Schedule 6 and one or more Codes of Practice for the purpose of compliance with other relevant duties including, but not limited to, those relating to the offences set out in Schedule 7. The Codes should describe with sufficient clarity and detail the measures we think would be proportionate to recommend for services to be compliant with their legal duties. In this consultation, we have developed measures for compliance with: the illegal content safety duties contained in the Act (sections 10 and 24), content reporting, so far as they relate to illegal content, (sections 20 and 31), and complaints procedures (sections 21 and 32).
- 11.4 We considered that physically separate Code documents for each of these would be repetitive and potentially confusing for stakeholders. Therefore, we have produced one document which sets out the three Codes covering all kinds of illegal harm. In this consultation document, we set out clearly where measures relate to CSEA content, terror content and/or other duties; and in our Codes themselves, we specify which measures are part of all three Codes, and which are part of a subset.
- 11.5 In designing our Codes, the Act requires us to have regard to several principles and objectives, and we must also consider our duties under the Communications Act 2003, the Equality Act 2010, the Northern Ireland Act 1998, the Human Rights Act 1998 and the Welsh language compliance notice applicable to Ofcom.<sup>4</sup> We provide further consideration of these in relation to each measure we recommend in the following chapters of this Volume (as well as in Chapter 8), in Chapter 24, where we set out how we have met the statutory tests laid out in Schedule 4 of the Act and in our Equality Impact Assessment and Welsh Language Assessment, which can be found in full in Annex 13.
- 11.6 The Codes relate to the design and operation of services in the UK or as it affects UK users of the service but apply to providers of such services from outside the UK.
- 11.7 Services that choose to implement the measures we recommended in our Codes of Practice will be treated as complying with the relevant duty. This means that Ofcom will not take enforcement action against them for breach of that duty if those measures have been

---

<sup>4</sup> [Ofcom Compliance Notice - Section 44 of the Welsh Language Measure \(Wales\) 2011.](#)

implemented. Service providers may seek to comply with a relevant duty in another way, but the Act provides that, in doing so, they must have regard to the importance of protecting users' right to freedom of expression within the law, and to the importance of protecting users from breaches of relevant privacy laws. Where providers do take alternative measures, they must keep a record of what they have done and explain how they think the relevant safety duties have been met. We talk more about our approach to enforcement in our guidance in Annex 11 and the requirement to keep records in Annex 6.

## Process and next steps

- 11.8 Once the Consultation period closes, we will consider and take into account responses and evidence received in order to prepare the final regulatory documents. After finalising and publishing our regulatory documents and conclusions, Ofcom must submit a Statement to the Secretary of State, who may set out further requirements (directions) for Ofcom in relation to our Codes where there are exceptional reasons relating to public health, national security, public safety, or relations with a government outside the United Kingdom. Otherwise, the Codes will be laid in Parliament. Unless either House of Parliament resolves not to approve the Codes within 40 days of them being laid, Ofcom will issue the Codes and they will come into force along with the duties to which they relate, 21 days later. In our draft enforcement guidance, we discuss implementation of the Codes and provide non-binding guidance on how quickly we might expect services to adopt the measures we recommend.
- 11.9 We will keep our Codes of Practice under review. Over time, we anticipate making further updates to our Codes through a process of iteration as our evidence base evolves. Updates to the Codes, other than minor amendments, will follow a similar Parliamentary procedure.

## Our key proposed recommendations

---

### Core illegal content Code measures

- 11.10 The majority of our proposed measures are designed to mitigate the risk of multiple kinds of illegal harm. This is in line with how we understand services themselves think about trust and safety measures. These include:
- a) Governance and accountability arrangements around the management of online safety risks, including senior management visibility of and accountability for key risks. We propose to recommend that all services establish clear accountability for compliance with their illegal content safety duty, complaints and risk assessment obligations, with additional expectations on large and multi-risk services.
  - b) Content moderation and search moderation measures focused on ensuring that regulated services comply with their duties around taking down or deprioritising illegal content; for some services it is proportionate to establish clear and appropriate internal policies, provide adequate content moderation resourcing in line with these policies, and provide up-to-date training materials to moderators.
  - c) Reporting and complaints recommendations for all services to make procedures easy for people to use, to allow users to provide extra information about their complaints (for example, to explain why content amounts to harassment), and to secure that they take appropriate action in response to complaints.

- d) Terms of service and publicly available statements providing clear and accessible information about these processes.
  - e) Large general search services handling complaints about predictive search recommendations.
  - f) Multi-risk services undertaking on-platform tests of recommender systems collect safety metrics and test for safety outcomes.
- 11.11 In addition, alongside our proposed measures tackling multiple kinds of illegal harm, we are proposing various measures targeted at specific kinds of illegal harm. These include measures targeted at CSEA, terror and fraud offences:
- a) U2U services use tools for detecting previously-identified child sexual abuse material (CSAM), specifically hash-matching and URL detection. We recommend this for services which assess themselves as being at medium to high risks of CSAM in their risk assessments and have a certain reach depending on certain functionalities. On general search services, we have recommended they use tools for URL detection and the use of warnings to add friction to potential perpetrators searching for CSAM.
  - b) For U2U services that have risks of grooming, specific changes should be made to default setting to protect for under 18s, and supportive information is provided to children using a service in a timely and accessible manner.
  - c) Large U2U services that have relevant risks should enable users to block and mute others and disable comments to mitigate risks including hate, harassment and encouraging suicide.
  - d) All U2U services block accounts being run by or on behalf of proscribed organisations.
  - e) Large U2U services that are risky for fraud implement keyword search tools to help them identify suspected illegal content relating to the sale of articles for use in frauds and establish dedicated fraud reporting channels to benefit from third party expertise in tackling online fraud.
  - f) Large services that have verification schemes and risk of fraud or foreign interference offences should establish clear internal policies for such verification schemes and improved public transparency for users.
  - g) Large general search services provide supportive information to users who search for suicide content.
  - h) Large general search services should provide content warnings and support resources to users who are clearly searching for CSAM.
- 11.12 Complete tables setting out the measures we propose, and who they apply to, are set out at the beginning of Chapter 23.

## **Building on our first Codes**

- 11.13 Our Codes will evolve over time as we learn more and as the sector develops. As set out in Ofcom’s approach to implementing the Online Safety Act, our aim with this first iteration of Codes is to capture existing good practice within industry, set clear expectations, and work to raise standards of user protection, especially for services whose existing systems are patchy or inadequate.

- 11.14 Recognising that we are developing a new and novel set of regulations for a sector without previous direct regulation of this kind, and that our existing evidence base is currently limited in some areas, these first Codes represent a basis on which to build, through both subsequent iterations of our Codes and our upcoming consultation on the Protection of Children. In this vein, our first proposed Codes include measures aimed at proper governance and accountability for online safety, which are aimed at embedding a culture of safety into organisational design and iterating and improving upon safety systems and processes over time.
- 11.15 In particular, there are certain areas where we are not yet making proposals. At this stage we are inviting respondents to share any further information they may hold in relation to the following:
- a) Our proposed recommendation around strikes and blocking in this consultation relates to proscribed groups. We are inviting further evidence from stakeholders to be able to explore broadening this in future work; in particular, we are aiming to explore a recommendation around user blocking relating to CSAM early next year. We are particularly interested in the human rights implications, how services manage the risk of false positives and evidence as to the effectiveness of such measures.
  - b) In Chapter 14, where we outline Automated Content Moderation (ACM) proposals for U2U services, we propose not to recommend at this time measures to detect terrorism content through hash matching and URL detection. Nevertheless, we recognise the potential importance of ACM in reducing the risk of terror-related online harms and are therefore using this consultation as an opportunity to gather further evidence.
  - c) We recognise that identifying previously unknown content is an important part of many services' processes for detecting and removing illegal content. We do not yet have the evidence base to set out clear proposals regarding the deployment of technologies such as machine learning or artificial intelligence to detect previously unknown content at this time. As our knowledge base develops, we will consider whether to include other recommendations on automated content classification in future iterations of our Codes.
  - d) We will continue to consider the use of 'trusted flagger' arrangements through dedicated reporting channels (DRCs) across all kinds of illegal harm. We are seeking further evidence from relevant stakeholders, particularly around the costs and impacts of such arrangements in specific harms areas so that we can consider further recommendations concerning DRCs in future iterations of our Codes.
- 11.16 Many of the measures we propose are for large services. This is often because we do not yet have enough information on the potential costs and benefits to know whether the measures are proportionate for smaller services at this point. As our understanding develops, it may be appropriate in future iterations of the Codes to expand the range of services for which some measures are recommended.
- 11.17 We note that services should be continuously monitoring and assessing risk through their risk assessment process. Reflecting current practice, many services may adopt further measures beyond those set out in the Codes to further protect users against sources of risk that they identify in their risk assessment. This may be particularly relevant for the largest



and riskiest services, with whom Ofcom is likely to have the closest supervisory relationships.<sup>5</sup>

- 11.18 We have also considered likely or confirmed regulatory provisions relating to online safety in other jurisdictions when developing our proposed measures. We believe there to be a good degree of alignment between our recommendations and those in other regulatory jurisdictions, including Australia and the EU, and that this is important in promoting compliance and minimising undue burdens on businesses. We are aware of other Codes or Code-like instruments being prepared or entering into force in other countries and will continue to monitor these developments and work constructively with other regulators, as set out in our approach to Ofcom’s approach to implementing the Online Safety Act.

## Our approach to developing recommended measures

---

### Our approach to our assessments

- 11.19 The Act requires us to take into account several principles (Schedule 4), which include:
- a) Consideration as to the appropriateness of the measures in Codes to providers of different kinds, sizes and capacities, including that measures we recommend in Codes are proportionate and technically feasible for different providers;
  - b) That providers must be able to understand which measures are relevant to their service, and that these measures are sufficiently clear and detailed for providers to understand what they entail in practice;
  - c) That the measures in the Codes are proportionate to Ofcom’s assessment of the risk of harm presented by services of that kind and size.
- 11.20 We must also ensure that measures described in Codes are compatible with pursuit of the online safety objectives listed in Schedule 4 and set out in full in Chapter 24 of this Consultation, which cover, in particular: the needs of different kinds of users and the overall user base; effectiveness and proportionality; access controls; and user protection in relation to algorithms, functionalities and (for U2U services) other features of the service and (for search services) the indexing, organisation and presentation of search results.
- 11.21 Schedule 4 also requires us to include measures in the Codes in each of the categories of measures contained within services’ safety duties in clauses 10(4) and 27(4). Under the Act, we are also required to carry out impact assessments when preparing a Code of Practice or amendment to a Code of Practice, including an assessment of the impact on small and micro businesses. We consider the Schedule 4 requirements in further detail in Chapter 24.
- 11.22 In line with these requirements, principles and objectives, to include a measure in the Codes, we need to assess that the measure is proportionate (with reference to both the risk presented by a service, and its size, kind and capacity) and does not unduly interfere with users’ rights to freedom of expression and privacy.
- 11.23 In this volume, we set out the main measures we have considered and our assessment of each. We have developed these proposals based on evidence drawn from stakeholder submissions to our 2022 Illegal Harms Call for Evidence, third parties and our own research.

---

<sup>5</sup> To better understand our likely approach to Supervision, please see Chapter 30.

This includes drawing on our assessment of the risks of illegal harms as set out in the register of risk in Chapter 6.

- 11.24 Where the duties in the Act are explicit that a particular measure is required, then we consider the measure in our Code to be proportionate. For example, under section 10(5) of the Act, all services must ensure that they have Terms of Service which are inclusive of certain information. The Act requires services which do not currently have such Terms of Service to develop them; the benefits or costs of doing so are not a matter in relation to which we exercise any discretion in our Codes of Practice.
- 11.25 In contrast, where there are material choices over what we might recommend in our Codes, and there are clear questions over what it is proportionate to do, and for which services, in order to comply with their duties, we have considered the costs and benefits of the proposed measures in some detail.
- 11.26 We have used proportionality as a key yardstick with which to decide whether to propose certain measures, and the final shape of those measures. To assess whether the proposals are proportionate, we considered the costs and benefits of different options, including how this might vary across services. As well as consideration of financial costs, the potential impacts on users – including both potential harm reduction and their human rights – was a central part of this assessment.
- 11.27 It is difficult to quantify costs and benefits of measures for a range of reasons, including:
- a) some of the costs and benefits are intangible in nature and difficult to quantify;
  - b) the broad scope of the regime means it is not possible to identify with confidence how many services are in scope, nor give a fully comprehensive description of the range of kinds of service in scope;
  - c) the diversity of services in scope means that costs will vary significantly between services depending on the characteristics of the service and specific choices they make about implementation.
- 11.28 We have sought to quantify costs and benefits where we can, but in recognition of the above, there are limits to the extent to which we have been able to quantify. In some cases, we have done a qualitative assessment of costs and benefits rather than a quantitative or monetised analysis. In many instances, we allow some flexibility in how services can practically implement our recommendations, to ensure these are appropriate and proportionate to their circumstances. In those areas where we are proposing to be more prescriptive around the details of practical implementation, our assessment and discussion of cost is also more detailed.
- 11.29 We have made some assumptions on costs that apply to multiple measures, such as salary assumptions for types of staff. We set these out at the beginning of Annex 14.
- 11.30 Where we are proposing to recommend a measure, we explain our rationale and supporting evidence. The core parts of our impact assessment cover:
- a) **Risk of harm:** except where the measure we are considering follows directly from the Act, we set out the risk of harm that we are seeking to address through our proposed measures.
  - b) **Likely efficacy and benefits** of the proposed measure in terms of reducing risk/harm, including any evidence derived from its current use by services of different kinds and

sizes. This evidence is important in setting out why the measure is proportionate when balanced against the potential costs and any human rights impacts.

- c) **Cost:** we consider costs broadly, including direct costs to services of implementing measures, and indirect costs such as loss of revenue. This analysis is important in considering the capacity of services of different kinds and sizes to adopt our proposed measures, particularly small businesses and micro businesses. We have quantified the potential costs where we can. Where the information is not available, we have instead described what we believe the nature of the costs may be, recognising that they are likely to vary widely by service. Our analysis has focused on costs on a 'per service' basis, rather than total cost to the sector, because this is often better for testing the proportionality of the proposed measures and can allow us to identify differences between services and because of uncertainty on the number of services in scope. In some cases, services may already have the same or similar measures we are proposing in place. Our analysis focuses on what the costs would be for those services that are **not** currently undertaking the measures, which is the more challenging test for whether the measures are proportionate. To the extent some services are already undertaking measures, and plan to continue doing so, then they would not face additional costs.
- d) **Human rights impacts:** any potential impacts on rights to freedom of expression, freedom of association and privacy.

- 11.31 This assessment can vary for different types of service. A key part of our assessment of proportionality for many measures is therefore which services they might be proportionate for. We discuss this in the next sub-section.
- 11.32 The level of detail and complexity in the comparison of costs and benefits is greater for some measures than others. This sometimes reflects the availability of information. It can also reflect where a more detailed assessment is more likely to impact our recommendations, and when it can affect which services we recommend measures for. This is especially the case for some of the measures we recommend to reduce grooming and the hash matching measure we recommend to reduce CSAM, where we carefully consider whether to recommend the measures for smaller services.
- 11.33 Our impact assessments have focused on considering individual measures on their own merits, but we recognise that measures would be applied in combination. It is possible therefore that our assessment may overstate the costs and benefits in some instances, particularly where implementing one measure would reduce the costs of implementing further measures or where measures are seeking to reduce the same harm. We have sought to consider these points when assessing options but have not formally assessed combinations of measures due to the complexities this presents. We set out our assessment of the impact of the cumulative set of measures that apply to different types of service in Chapter 23.
- 11.34 We have also considered the **equality impacts** of our proposed measures, setting out our understanding of any particular impacts on protected groups in the UK, which can be found in Annex 13. Generally speaking, we expect that our proposals are likely to promote equality.
- 11.35 Where relevant, we have also considered likely or confirmed regulatory provisions relating to online safety in other jurisdictions when developing our proposed measures. Services will also need to ensure that they comply with data protection law and, where relevant, the

Privacy and Electronic Communications Regulations (PECR).<sup>6</sup> Users' rights to data protection are covered by UK GDPR and the Data Protection Act 2018 which are enforced by the Information Commissioner's Office (ICO).<sup>7</sup> The ICO has a range of data protection and PECR compliance guidance which we encourage services to consult.<sup>8</sup> Services likely to be accessed by children should also ensure they conform with the ICO's Children's Code.<sup>9</sup>

- 11.36 The measures we are proposing may have different impacts on services of different kinds and sizes. As a result, some of our recommendations apply to all U2U and search services; others may apply differently to different kinds of services; and some apply only to specific services and for specific risks.

## Which services do the Codes measures apply to

- 11.37 We propose recommending a small number of measures for all U2U services, and a similar set of measures for all search services. Most of these proposed measures relate fairly directly to explicit duties in the Act. For example, this applies to measures related to Terms of Service.
- 11.38 Beyond this minimum set of measures, the measures we recommend for any service depend on combinations of the following factors:
- a) The risks the service has found in its latest illegal content risk assessment.
  - b) The size of the service.
  - c) Whether the service is a U2U service or a search service. With search services, we also distinguish between vertical search services and general search services, and propose more measures for general search services than vertical search services.
  - d) The functionality or user base of the service. For example, some of the measures to reduce grooming on U2U services only apply if there are children using the service and if the service has relevant functionalities including user connections or direct messaging.
  - e) Whether the service is a U2U service or a search service. With search services, we also distinguish between vertical search services and general search services, and propose additional measures for general search services.
- 11.39 Below we expand on our general approach to the first three of these, namely how a service's illegal content risk assessment affects the measures we propose to recommend, our proposed approach to varying measures by the size of the service, and how we are thinking about the types of search service.
- 11.40 In our draft Codes<sup>10</sup>, we set out within each proposed recommended measure which services they apply to, as well as summarising this at the beginning of the document, ahead of the detail of the measures themselves.

## Application of harm specific measures to risky services

- 11.41 Some of the measures we recommend are targeted at addressing the risk of specific kinds of illegal harm, such as our measures relating to CSAM, grooming, fraud and foreign

---

<sup>6</sup> The Privacy and Electronic Communications Regulations 2003.

<sup>7</sup> The Data Protection Act 2018.

<sup>8</sup> For further guidance, see please see the [ICO for organisations](#).

<sup>9</sup> ICO, [Age Appropriate Design Code](#) (which we refer to as the 'Children's Code'), 2022.

<sup>10</sup> See Annexes 7 and 8 for our draft proposed Codes of Practice.

interference offences. Whether these harm-specific measures are proportionate can depend on how risky a service is for the relevant kinds of harm. Typically, our recommendations only apply to services that have identified a medium or high risk of relevant kind of harms in their latest illegal content risk assessment.

- 11.42 The draft Services Risk Assessment Guidance we have produced sets out how we expect services to assess their risk level, with reference to Ofcom's Risk Profiles, and determine whether they are medium or high risk for different kinds of priority offences. We recommend that services refer to our guidance to satisfy themselves that they are meeting their duty to conduct a suitable and sufficient risk assessment, and to understand which Codes measures would likely apply to them given the risks that they have identified.

### We propose applying some measures to 'multi-risk' services

- 11.43 As well as harm-specific measures relating to particular kinds of illegal offences, we also propose recommending some measures targeting all harms. Rather than being aimed at specific kinds of illegal harm, these are aimed at illegal harms generally. In particular, we propose recommending some measures that should address all kinds of illegal harm – examples are governance and some of our proposed content moderation measures.
- 11.44 We intend these measures to apply to services that face significant risks for illegal harms in general. There is a question over what it means for a service to have such risks. One option would be to recommend these measures to services that have identified as medium or high risk of at least one kind of illegal harm. However, where services only identify a risk of a single kind of illegal harm, the benefits of these measures to address all harms will be lower. This is partly because if services have only identified a single area of risk, the extent of harm will tend to be lower compared to if they have identified a range of kinds of offence where they are high risk. It is also partly because many of these measures are about enabling services to have a good understanding of their risks and of the content moderation policies needed to address those risks. If a service was only of medium or high risk for a single kind of illegal harm, the risk is more likely to be well understood across the organisation, such as the risk of fraud for some marketplace services. This tends to mean the benefits of these measures in terms of improving understanding and consistency of approach are smaller than if there were multiple areas of risk. The case for the measures to address all harms being proportionate therefore tends to be stronger if we only apply them to services that have identified multiple kinds of illegal harm.
- 11.45 On the other hand, if we set the threshold for when the measures are recommended at a high level, for example if a service were risky for many kinds of illegal harm, then the overall benefits from measures would be smaller, as they would be recommended for fewer services.
- 11.46 On balance, we **propose some measures for services that identify as medium or high risk for at least two different kinds of illegal harm** in their latest illegal harms risk assessment. We refer to such services as **multi-risk**.

### We consider small and micro businesses specifically

- 11.47 We are required under the Act to consider the impact of our proposed measures on small and micro businesses, in particular. We have used what we understand to be the definitions across many Government bodies for defining these businesses, based on numbers of full-time employees. The definitions are: businesses that employ 10-49 full-time employees for

small businesses, and between 1-9 full-time employees for micro businesses.<sup>11</sup> We understand these definitions to mean employees located in the UK or outside the UK. Even measures which might be considered to have only ‘low’ to ‘moderate’ costs may have a significant cost impact on micro businesses.

- 11.48 As described below, we propose to recommend some measures only for large services. This means that we can avoid unduly burdening small and micro businesses. However, we also want to make clear that our approach does not mean that such services will always be exempted from measures which result in a significant financial burden. In some circumstances we would expect that any negative financial impact would be justified, having regard to the risk of harm identified in a risk assessment, necessitating changes to protect users.<sup>12</sup> We do therefore recommend some measures to small and micro businesses that might be costly to implement, where those businesses create significant risks for users.
- 11.49 As well as considering each proposed measure individually, we summarise the set of measures we propose for small and micro businesses in Chapter 23 and consider the cumulative impact on those services.

## We propose more measures for large services

- 11.50 We propose that many Codes measures are only recommended for ‘large services’. We often apply these recommendations to services that have also identified high or medium risk of certain illegal harms.
- 11.51 **We propose to define ‘large service’ as a service with a number of monthly UK users that exceeds 7 million.** This is roughly 10% of the UK population, and broadly equivalent to ‘services with a large user base’ in the Register of Risks. We note that this approach of taking user base as a proxy for the size of service is similar to that adopted by the EU in the DSA.<sup>13</sup> We consider it important to broadly align our approach to determining larger services with other international regimes where possible, to reduce the potential burden of regulatory compliance for services. Data from Ipsos iris found 109 online brands were each visited by over 7 million UK individuals aged 15+ on smartphones, tablets or computers in January 2023.<sup>14</sup> Only some fraction of these 109 brands would have regulated services and be in-scope for the Act.
- 11.52 Part of the reason for recommending some measures only for large services is that the benefits of measures are likely to be greater for such services, because more users will be protected by the measure.

---

<sup>11</sup> We appreciate that not all Government bodies use exactly the same definitions. For example, some also refer to revenue and assets. The definition we propose is consistent with that used by the Regulatory Policy Committee. It would not make a material difference to our impact assessment if another common definition of small and micro business (such as that consistent with the Companies Act 2006) were used instead. Source: Regulatory Policy Committee, 2019. [Small and Micro Business Assessments: guidance for departments, with case history examples, August 2019 \[accessed 11 September\]](#).

<sup>12</sup> This includes some circumstances where a service may need to close if it is unable to introduce a measure which is required to mitigate a significant risk of harm.

<sup>13</sup> The DSA classifies platforms or search engines as very large online platforms (VLOPs) or very large online search engines (VLOSEs) if they have more than 45 million users per month in the EU, a number equivalent to 10% of the EU population.

<sup>14</sup> We note Ipsos defines an online brand as consisting of its applications and websites. Source: Ipsos, [Ipsos Online Audience Measurement Service](#), January 2023, age: 15+, UK.

- 11.53 Another reason for restricting measures to large services is that for some proposed measures the costs for services may be significant, and those costs could have a material effect on the operation of non-large services.<sup>15</sup> This is part of how we have taken into account the capacity of services when considering the proportionality of measures. We assume that in general large services have more resources available to undertake measures than smaller services. We consider it can be prudent to exempt smaller services from incurring those costs (where appropriate provided they are not high risk), as there will often be significant uncertainty in any assessment of benefits and costs, and we want to reduce the possibility of imposing financially damaging costs on businesses when the magnitude of benefits expected to result from the measure is uncertain.
- 11.54 Also, in some cases, imposing costly measures on smaller services could reduce their ability to operate and compete effectively in certain online markets. A lack of competition and innovation can be very costly for society and needs to be considered against the scale of potential benefits from any measure.
- 11.55 While we assume that, in general, large services have more resources available, it is possible that a service has a large user base, while still having limited capacity to undertake measures. If we define large purely in terms of user base, there is therefore a risk that a service classified as large is actually a low-capacity service, and does not have access to significant resources.
- 11.56 We could supplement the definition of large so that, as well as user base, it includes a condition relating to the services' access to resources. The capacity to undertake measures could relate to either access to financial resources or having technical expertise. However, with time, we consider that the lack of technical expertise could be addressed by hiring or contracting for the necessary expertise, provided a service had access to sufficient financial resources to do this. We therefore consider that it is ultimately access to financial resources that is most important for determining the capacity of a service to undertake a measure. Profit or revenue could be proxies for access to financial resources. We could therefore supplement our definition of large with a specified amount of profit or revenue. For services serving many countries, we would want to consider finances related to the UK only. Using revenue would be simpler than profit as it would require the allocation of revenue without also requiring the allocation of costs. We could set a revenue limit of £10 to £50 million UK-related revenues a year. This would have the advantage of excluding services that have low resources despite having a large user base.
- 11.57 However, revenue (and profit) may not always accurately capture whether a service has access to resources. For example, a service could have good access to capital because of its anticipated future revenue stream due to its high user base, even though it had not yet fully monetised that user base.
- 11.58 Another option would be to add a condition to the definition of 'large' that referred to the number of employees of the provider of the service. With the definitions we propose, it is theoretically possible for a service to be a large service (having an average reach of over 7 million UK users) and for it to be provided by a small business (less than 50 employees). However, we do not anticipate any services currently being in this category. For a service to serve 7 million UK users, the provider is likely to need more than 50 employees. However,

---

<sup>15</sup> This is particularly likely to be the case when the measures require large one-off costs that do not vary in magnitude with the size of the service.



we cannot rule out the possibility that an exceptional service could have large reach while being provided by a business with few employees. An advantage of defining ‘large’ so that it also incorporates the number of employees is that it would make this impossible. This could be an alternative to, or as well as, a condition relating to revenue. However, our provisional view is that this is unnecessary. In the unlikely event there were a service with a reach of over 7 million UK users, the risk to those users and the likely access to capital of a service with that reach means that it could be proportionate for it to adopt the measures that other ‘large’ services should adopt.

- 11.59 As with the other possibilities for measuring size (such as user base, revenue and profit), the number of employees that a service has does not perfectly predict whether that service can access resources or not. For example, it is possible that some businesses with few employees are able to access significant financial resources if backers are confident in its future ability to earn revenues.
- 11.60 On balance, we propose to keep the definition of ‘large’ simple and relate it only to whether a service has, on average, more than 7 million monthly UK users. This avoids the need for services to determine what share of their revenues relates to the UK. This also will ensure that all services are captured when they have high reach as these are the services where the benefits of applying measures are likely to be greatest. It is also broadly consistent with the approach taken by the EU in the DSA.
- 11.61 We recognise that providers will differ in how they measure users for this purpose. The way users are defined in the Act is discussed in chapter Overview of Regulated Services from paragraph 3.12.
- 11.62 We think it’s important that there is some uniformity in the way that a providers calculate their average number of monthly users. We consider the period of twelve calendar months leading up to the point at which a provider makes its calculation to be a large enough sample against which to take a meaningful average figure. To prevent providers from having to deal with periods that straddle calendar months (which may be difficult to determine), we propose that providers calculate by reference to the twelve-calendar month period ending with the month before the month in respect of which the provider is making its calculation.
- 11.63 We also want to provide certainty to providers whose monthly figures may come close to the 7 million mark, and avoid doubt as to their status where, in a given month, the average monthly figure may fall below the 7 million mark. We therefore propose that once a service calculates that it counts as a large service (in accordance with the principles described above), it is to be treated as such for the purposes of adhering to measures in the Code unless the average number of its monthly UK users does not exceed 7 million for an uninterrupted period of six months or more. In those circumstances, the provider should not treat itself as a large service (and, accordingly, it shouldn’t be treated as such by Ofcom). In practice, services are therefore encouraged to keep track of how their average user figure fluctuates month by month.
- 11.64 We return to the definition of large services in Chapter 23 when we draw together the measures we proposed to apply to large services.

## Taxonomy of search services

- 11.65 When we describe the measures we recommend for search services, we distinguish between the following types of search services. These definitions are based on the definitions in the



Act, our understanding of how search services operate on a technical basis. The definition of ‘large’ services above also applies.

- a) **General search services:** General search services enable users to search the contents of the web by inputting search queries on any topic and returning results. There are two types of general search service:
- i) **General search services which only rely on their own indexing:** They work by using crawlers (also called bots) to find content across the web (‘crawling’); building an index of URLs by validating and storing the content found in a database (‘indexing’); and using algorithms – for example, Google’s PageRank – to rank the content based on relevance to the search query (‘ranking’). Search services use many ranking signals, the details of which are proprietary and not publicly known.<sup>16</sup> There are a small number of large general search services that do their own crawling and indexing, which provide search results to downstream general search providers. There are also smaller search services which do their own indexing.<sup>17</sup>
  - ii) **Downstream general search services:** As a type of general search service, downstream general search services provide access to content from across the web, but they are distinct in that they obtain (or supplement) their search results from those general search services which rely solely on their own indexing. Depending on their contract with a general search service, downstream general search services may have limited control over how search services are displayed.<sup>18</sup> Downstream general services also often distinguish themselves from general search services by offering a social purpose (e.g. Ecosia), additional privacy (e.g. DuckDuckGo), or differentiated search features.
- b) **Vertical search services:** Also known as ‘speciality search engines’, these operate differently from general search services. Rather than crawling the web and indexing webpages, they present users with results only from selected websites with which they have a contract, and an API or equivalent technology<sup>19</sup> is used to return the relevant content to users. Common vertical search services include price comparison sites and job listing sites.

11.66 There is a lack of clear evidence to suggest that vertical search services play a significant role in the dissemination of priority illegal content or other illegal content.<sup>20</sup> We therefore propose to recommend fewer measures for vertical search services. This is particularly the case for when we have proposed measures specifically for ‘large’ services, where we have generally proposed excluding large vertical search services.

11.67 We recognise that downstream general search services may have limited control over the ranking of search content that might be accessed via their service, as it will depend on their

---

<sup>16</sup> The search engine index takes the output from the crawler and creates relevant data structures to support later searching within the search engine. The index can comprise document content, images, and metadata. An index will have many repeated refinement algorithms applied to increase its accuracy and relevance.

<sup>17</sup> Mojeek describes how it does not have a syndication agreement with a large general search services (and hence is not what we are calling a downstream general search service) .Source: Mojeek [response to the UK Government’s Online Safety Bill Impact Assessment](#), July 2021, page 2.

<sup>18</sup> In its advertising market study, the CMA said none of the contracts it had looked at allowed the downstream general search service to re-rank the search results they received from Google or Bing. Source: CMA, 2020. [Online platforms and digital advertising: Market study final report](#), Box 3.3 page 97 and paragraph 3.85.

<sup>19</sup> Application Programming Interface: a way for two or more computer programs to communicate with each other.

<sup>20</sup> Please see Volume 2: Chapter 6T (Search services), paragraph 6T.21

contract with a general search service. It may therefore be difficult for them to implement some measures directly. Nevertheless, the duties in the Act, and the measures we recommend services take to meet those duties, still apply to such services. If they do not wish, or it is not technically possible, for them to adopt the relevant measures themselves, they can secure by contract that the relevant duties are met.

## Structure of this volume

---

11.68 The following chapters in this volume are as set out below

- Governance and accountability (Search and U2U) [in volume 3].
- Content Moderation (U2U)
- Search Moderation (Search)
- Automated Content Moderation (U2U)
- Automated Moderation (Search)
- Reporting and Complaints (U2U and Search)
- Terms Of Service and Publicly Available Statements (U2U and Search)
- Default settings and user support (U2U)
- Recommender systems (U2U)
- Enhanced User Controls (U2U)
- User Access (U2U)
- Search features, functionality and user support (Search)
- Cumulative assessment
- Statutory tests

## 12. U2U content moderation

### What is this chapter about?

This chapter sets out our proposed recommendations regarding how services should set up their content moderation systems to meet their duties relating to illegal harms. It is important to make clear that, as the regulator, Ofcom will not take a view on individual pieces of online content. Rather, our regulatory approach is to ensure that services have the systems and processes in place to meet their duties.

### What are we proposing?

We are making the following proposals for all U2U services:

- **Have systems or processes designed to swiftly take down illegal content of which it is aware.**

We are making the following proposals for all multi-risk U2U services and all large U2U services:

- **Set and record internal content policies. These should set out rules, standards and guidelines about: what content is allowed and not allowed on the service, and how policies should be operationalised and enforced. In doing so, services should have regard to its risk assessment and signals of emerging illegal harm.**
- **Set and record performance targets for its content moderation functions and measure and monitor its performance against these targets.** These should include targets for both how quickly illegal content is removed and for the accuracy of content moderation decisions. When setting performance targets services should balance the need to take illegal content down swiftly against the need to make accurate moderation decisions.
- **Prepare and apply a policy about the prioritisation of content for review. This policy should have regard to at least the following factors: virality of content, potential severity of content, and the likelihood that content is illegal, including whether it has been flagged by a trusted flagger.**
- **Resource its content moderation function so as to give effect to its internal content policies and performance targets.** In doing so, it should have regard to the propensity for increases in demand for content moderation caused by external events. When deciding how to resource their functions services should consider the particular needs of its UK user base, in relation to languages.
- **Ensure people working in content moderation receive training and materials that enable them to moderate content effectively.**

### Why are we proposing this?

Content moderation is the practice of identifying and reviewing content to decide whether it should be permitted on a service. Effective content moderation systems or processes allow services to identify and remove illegal content swiftly, accurately and consistently. The available evidence shows that content moderation plays a hugely important role in combatting online harms and that services with ineffective content moderation functions pose an increased risk of harm to users.

Our analysis suggests that harm to users will be reduced where services set content policies, resource and train their content moderation teams adequately and take into account the likely severity of content and the risk it will go viral when deciding what potentially harmful content to prioritise for review. Given the diverse range of services 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 proposals about content moderation, we are therefore consulting on 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 onerous 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 services to carry them out in a way which is cost-effective and proportionate for them.

We recognise that services often use a combination of automated tools and human review to moderate content. The proposals in this chapter are not prescriptive about the balance services should strike between human and automated review of content and would not require services to use automated tools to review content. Given the important implications they would have for privacy rights, where we have made specific recommendations about automated review of content we consider these separately and in more detail in a later chapter.

### What input do we want from stakeholders?

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

## Introduction

---

- 12.1 Content moderation is when a service reviews content to decide whether it is permitted on its service. It can be done by humans, automatically or by a combination of the two.<sup>21</sup> It includes the identification and assessment of content and any actions taken on content.<sup>22</sup> This can include rules imposed on content, the human labour and technologies required, and the institutional mechanisms of enforcement and appeals that support it.<sup>23</sup> It should be noted that ‘appeals’ are considered in Chapter 16. In addition, we set out specific recommendations in relation to automated content moderation for U2U services in Chapter 14.
- 12.2 There is no ‘one-size-fits-all’ approach to content moderation. Content moderation systems and processes differ from service to service and are designed to meet specific needs and contexts.<sup>24</sup> We know that content moderation systems, particularly those deployed across a very large user base, cannot provide a guarantee that users will not encounter any illegal content (and are often designed around reducing instances, rather than complete

---

<sup>21</sup> Encyclopedia of Big Data, 2017. [Content Moderation](#). [accessed 2 August 2023].

<sup>22</sup> Gillespie, T., and Aufderheide, P., 2020. *Expanding the debate about content moderation: scholarly research agendas for the coming policy debates*. In: *Internet Policy Review*, 9(4).

<sup>23</sup> Tarleton Gillespie, T., and Aufderheide, P., 2020. *Expanding the debate about content moderation: scholarly research agendas for the coming policy debates*. In: *Internet Policy Review*, 9(4).

<sup>24</sup> Gillespie, T., and Aufderheide, P., 2020. *Expanding the debate about content moderation: scholarly research agendas for the coming policy debates*. In: *Internet Policy Review*, 9(4); The Center for Democracy & Technology (CDT), 2021. [Outside Looking In Approaches to Content Moderation in End-to-End Encrypted Systems](#). [accessed 2 August 2023].

prevention). However, well-designed and resourced content moderation systems and processes can significantly reduce risks and help protect users.

- 12.3 While specific content moderation requirements are likely to differ between services depending on a range of factors, we consider there to be certain core measures that will secure compliance with the safety duties.
- 12.4 We recognise that many services set up their content moderation systems and processes to deal with both illegal and legal content so as to align with their published terms of service. However, for this first phase of our work, the content moderation measures we are proposing will be focused on dealing with illegal content.

## Harms the measures seek to address

- 12.5 Under section 10 of the Act, regulated U2U services must 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 users from illegal content (section 10(2)(a), (b) and (c)).
- 12.6 Alongside this, services have a duty to have proportionate systems and processes to minimise the length of time for which any priority illegal content is present (section 10(3)(a)). Where the service is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way, services must swiftly take down such content (section 10(3)(b)).
- 12.7 As set out in Chapter 16, services also have a duty to take appropriate action in response to complaints about illegal content, and to handle appeals about action taken against content or a user because content is identified as illegal.
- 12.8 Compliance with these duties, in particular the duties to take down illegal content swiftly on becoming aware of it and to take appropriate action in response to complaints about illegal content, would be very difficult in practice absent some process for determining whether or not content ought to be taken down and implementing that decision as appropriate.
- 12.9 In practice, content moderation is used by services to address a wide variety of illegal harms, as well as legal content that does not comply with a service's own policies. The overall effect of having a content moderation process is to support compliance with legal obligations, to help keep users safe, and to maintain a trusted environment for other actors, including advertisers where relevant.<sup>25</sup>
- 12.10 Content moderation systems and processes typically rely on a service's content policies, which may include, though are not limited to, terms of service, community guidelines and other relevant documents, which form the basis for content moderation practices. Collectively, these documents tend to dictate how violative content will be moderated.<sup>26</sup> While content policies usually prohibit the posting of illegal content, due to the global

---

<sup>25</sup> Trust & Safety Professional Association, no date. [What Is Content Moderation?](#) [accessed 2 August 2023]; New America, 2019. *Everything in Moderation: An Analysis of How Internet Platforms are Using Artificial Intelligence to Moderate User-Generated Content*.

<sup>26</sup> Cambridge Consultants, 2019. [Use of AI in Content Moderation](#). [accessed 3 August 2023]; Google, 2020. [Information quality and content moderation](#). [accessed 3 August 2023]; 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 3 August 2023]; Ofcom, 2023. [Content moderation in user-to-user online services: An overview of processes and challenges](#). [accessed 25 September]; .

nature of many services, content policies do not necessarily closely reflect the requirements of any single legal system.<sup>27</sup>

- 12.11 Effectively enforced content moderation is one of the key ways in which services can reduce the risk of users encountering illegal content of all kinds. Conversely, a lack of effective and consistently applied content moderation processes can lead to an increased risk of illegal content and subsequent harm to users.
- 12.12 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.<sup>28</sup>
- 12.13 A report by CASM Technology and ISD found a major increase in the number of antisemitic posts coinciding with a reduction in content moderation staff at one social media service, saying the analysis demonstrates *“the broader and longer-term impact that platforms de-prioritising content moderation can have on the spread of online hate.”*<sup>29</sup> Similarly, in late 2022, the Anti-Defamation League (ADL), noted an increase in antisemitic content on the same service and a decrease in the moderation of antisemitic posts.<sup>30</sup>
- 12.14 Another 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.”*<sup>31</sup>
- 12.15 While these examples are not solely related to illegal content, we think that they demonstrate the risks associated with less effective systems of content moderation.
- 12.16 Ineffective or poorly resourced content moderation appears to have serious impacts on user safety across a wide range of illegal or potentially illegal harms. There have been several examples of online services’ content moderation systems failing to tackle illegal harm or being used to facilitate illegal offences.<sup>32</sup> Conversely, some of these examples also serve to

---

<sup>27</sup> 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 3 August 2023].

<sup>28</sup> The Institute for Strategic Dialogue, 2021. [Gaming and Extremism: The Extreme Right on Twitch](#). [accessed 3 August 2023].

<sup>29</sup> CASM Technology and ISD, 2023. [Antisemitism on Twitter Before and After Elon Musk's Acquisition](#). [accessed 3 August 2023].

<sup>30</sup> ADL, 2022. [Extremists, Far Right Figures Exploit Recent Changes to Twitter](#). [accessed 3 August 2023].

<sup>31</sup> HOPE not hate and the Antisemitism Policy Trust, 2021. [Antisemitism and Misogyny: Overlap and Interplay](#). [accessed 3 August 2023].

<sup>32</sup> House of Commons Home Affairs Committee, 2017. [Hate crime: abuse, hate and extremism online](#). [accessed 3 August 2023]; Counter Extremism Project, 2018. [OK Google, Show Me Extremism: Analysis of YouTube's Extremist Video Takedown Policy and Counter-Narrative Program](#). [accessed 3 August 2023]; BSR, 2018. [Human Rights Impact Assessment: Facebook in Myanmar](#). [accessed 3 August 2023]; Meta, 2018. [An Independent Assessment of the Human Rights Impact of Facebook in Myanmar](#). [accessed 3 August 2023]; Amnesty International, 2022. [Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya](#). [accessed 3 August 2023].

demonstrate increased user safety and a reduction in illegal (or harmful) content when investment is put into improving content moderation systems.<sup>33</sup>

- 12.17 The harms we consider in this section potentially arise on all U2U services, but to different degrees. Some services, for example low risk/smaller services, may not have very much content to moderate (e.g. because they receive few complaints, because proactive content detection technology is beyond their means,<sup>34</sup> or because their business model is such that there is little likelihood of users uploading any illegal content without the service knowing about it). By contrast, larger and higher risk services may face significant challenges in terms of the volumes and diverse nature of the content they need to moderate, giving risk to questions about how to prioritise content for review, achieve consistency, quality and timeliness of decision-making, and plan their deployment of moderation resourcing so as to secure that users are appropriately protected.

## Proposed approach

---

### How we have approached the provisions in Codes

- 12.18 In light of the analysis above, we consider that it is important to include recommendations about content moderation in our Codes of Practice.
- 12.19 We have considered three potential approaches to drafting these measures:
- **Approach 1** - specify in detail how services should configure their content moderation systems and processes;
  - **Approach 2** - specify in detail the outcomes content moderation systems and processes should achieve (i.e. setting detailed KPIs), but leave the design to services, or;
  - **Approach 3** - require services 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.
- 12.20 These are not mutually exclusive – it might be possible to take different approaches in different areas. Below, we explore the benefits and drawbacks of each option.

### Approach 1: Specify in detail how services should configure their content moderation systems and processes

- 12.21 This section examines whether Codes should specify in detail how services should configure their content moderation systems and processes. For example, Codes could specify in detail what resources services should have in place (e.g. number of content moderators, specific automated systems, etc), what specific training staff should receive, or what specific KPIs services should set and the thresholds these KPIs expected to achieve.

---

<sup>33</sup> Google, 2020. [Information quality and content moderation](#). [accessed 3 August 2023]; Reddit, 2022. [2022 Transparency Report](#). [accessed 3 August 2023]; Google, no date. [Featured policies: Violent extremism](#). [accessed 3 August 2023].

<sup>34</sup> Ofcom, 2023. [Content moderation in user-to-user online services: An overview of processes and challenges](#). [accessed 25 September 2023].



- 12.22 We know that many services use a hybrid approach to enforcing content policies, i.e. using both human and automated resources.<sup>35</sup> There is a high possibility that moderation decisions at small and low risk services would primarily be made manually by small teams of moderators, or even members of senior management (the latter with a vested interest in making good decisions), rather than by larger teams of content moderators. In its response to the 2022 Illegal Harms Call for Evidence, [CONFIDENTIAL X].<sup>36</sup> [CONFIDENTIAL X].<sup>37</sup> If they have automated technology at all it is likely to be trained by a third-party (i.e. ‘off-the-shelf’ tools), rather than bespoke and/or specially trained automated technology.<sup>38</sup>
- 12.23 In larger services, Trust and Safety teams may be typically responsible for ensuring that policies are enforced with the appropriate resource. Trust and Safety teams may work closely with product and engineering teams which are responsible for supporting the content moderation process by developing the required tools and infrastructure or iterating the design of services to reduce the load on content moderation teams.
- 12.24 There is less information on how smaller services establish and deploy their content moderation resource and there is no one way they do this, but evidence indicates these services may rely more heavily on human moderation, which may be supported by ‘off-the-shelf’ automated moderation tools.<sup>39</sup> For example, in response to the 2022 Illegal Harms Call for Evidence, Synthesia, an artificial intelligence (‘AI’) video platform, said it uses “a mixture of in-house human moderation, along with machine-based tools” to moderate content.<sup>40</sup>
- 12.25 Most services include humans in their content moderation systems and processes, with some services outsourcing human resource, notably content moderators, to third parties.<sup>41</sup> The wider content moderation ecosystem can include a range of other staff, including, but not limited to: Trust and Safety policy staff; quality assurance staff; subject matters experts; risk management staff; operations staff; engineers; and developers.<sup>42</sup>
- 12.26 Nevertheless, there is little available evidence on how services deploy this human resource across their content moderation systems to deal with illegal and/or harmful content. Where human reviewers are used, it is possible to have different teams for different types of harm, and/or different teams for different reporting channels (e.g. flags or reports from trusted flaggers could be channelled to different teams, or could be fed into one team).<sup>43</sup>
- 12.27 Automated content moderation (ACM) tools also make up a resource that can be deployed across systems to tackle illegal and/or harmful content. Due to the complexities of harms, and the intrinsic limitations of individual automated content moderation technologies, it is

---

<sup>35</sup> Cambridge Consultants, 2019. [Use of AI in Content Moderation](#). [accessed 3 August 2023].

<sup>36</sup> [CONFIDENTIAL X].

<sup>37</sup> [CONFIDENTIAL X].

<sup>38</sup> New America, 2019. [Everything in Moderation – Introduction](#). [accessed 22 August 2023]; New America, 2019. [Everything in Moderation - The Limitations of Automated Tools in Content Moderation](#). [accessed 22 August 2023]; Gillespie, T., 2020. [Content moderation, AI, and the question of scale](#). [accessed 22 August 2023]; Chowdhury, N., 2022. [Automated Content Moderation: A Primer](#). [accessed 3 August 2023].

<sup>39</sup> New America, 2019. [Everything in Moderation – Introduction](#). [accessed 22 August 2023].

<sup>40</sup> Synthesia, 2022. [Synthesia response to the Ofcom 2022 Illegal Harms Call for Evidence](#): First phase of online safety regulation

<sup>41</sup> New America, 2019. [Everything in Moderation – Introduction](#). [accessed 22 August 2023].

<sup>42</sup> Trust and Safety Professional Association, no date. [Setting Up Content Moderation Teams](#). [accessed 3 August 2023].

<sup>43</sup> Trust and Safety Professional Association, no date. [Setting Up Content Moderation Teams](#). [accessed 3 August 2023].



often the case that services will use several automated tools in conjunction, layering one measure on top of another, as well as other signals, to assess with sufficient confidence whether a piece of content is violative or illegal and should be removed.

12.28 ACM tools are developed to identify illegal (and/or harmful/violative) content by following rules (typically those set out in Community Guidelines and other content policies).<sup>44</sup> Many services deploy such tools to enforce their content policies.<sup>45</sup> Grindr, for example, states that it uses ‘proprietary technological tools’ to help it proactively flag illicit content<sup>46</sup>, while Meta states it is increasingly using an ‘automation-first approach’ to content moderation to review more content across all types of policy violations.<sup>47</sup> [CONFIDENTIAL X].<sup>48</sup> While we know some services use various forms of automated content moderation (ACM) tools to identify content for moderation, we currently have limited information about most of these.

12.29 There is significant diversity and innovation in content moderation processes.

12.30 The benefits to adopting approach 1 for our Codes, and specifying what services’ content moderation processes should look like, would be:

- **Raise standards** - A high level of specificity around how services should configure their content moderation systems and processes could potentially raise the standard of content moderation systems and processes. For example, telling services what specific training they should give their staff and how to carry out this training could ensure staff involved in content moderation are better equipped to carry out their job, thus increasing standards.
- **Ease of compliance** – A high level of specificity around how services should configure their content moderation systems and processes could create clarity for services on what they should do, in turn allowing greater certainty for services that they are compliant with the Codes and making it more likely that they will do enough to protect users appropriately.

12.31 The drawbacks to this approach would be:

- **Diversity of services** - There is no ‘one-size-fits-all’ approach to content moderation. Content moderation systems and processes differ from service to service and are designed to meet specific needs and contexts.<sup>49</sup> What is appropriate and effective will vary from service to service depending on its characteristics. If Codes were to impose specific requirements on all services as to how they should configure every aspect of their content moderation systems and

---

<sup>44</sup> Cambridge Consultants, 2019. [Use of AI in Content Moderation](#). [accessed 3 August 2023].

<sup>45</sup> Google, 2020. [Information quality and content moderation](#). [accessed 3 August 2023]; Meta, 2020. [How We Review Content](#). [accessed 3 August 2023]; M. Singhal, et al., 2023. SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. In: 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P); Bumble, no date. [Guidelines](#). [accessed 3 August 2023]; Grindr, no date. [How Grindr moderates content and profiles](#). [accessed 3 August 2023]; Snap, no date. [How We Rank Content on Discover](#). [accessed 3 August 2023].

<sup>46</sup> Grindr, no date. [How Grindr moderates content and profiles](#). [accessed 3 August 2023].

<sup>47</sup> Meta, 2020. [How We Review Content](#). [accessed 3 August 2023].

<sup>48</sup> [CONFIDENTIAL X].

<sup>49</sup> Tarleton, G., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernandez, A., Roberts, S. T., Sinnreich, A., and West, S. M., 2020. [Expanding the debate about content moderation: scholarly research agendas for the coming policy debates](#). [accessed 3 August 2023]; The Center for Democracy & Technology (CDT), 2021. [Outside Looking In Approaches to Content Moderation in End-to-End Encrypted Systems](#). [accessed 3 August 2023].

processes, the requirements could be irrelevant or unapplicable to some services or could even be detrimental to user safety if they are not the most appropriate/effective interventions for particular services. This approach would also ignore the fact that there can be multiple approaches used to configure systems and processes to improve user safety. It could inappropriately deter innovation.

- **Lack of evidence** – At this stage, there is a lack of evidence and little consensus on the most effective way to configure content moderation systems and processes generally, although we consider exceptions to this general position in Chapter 14.
- **Impacts on smaller and diverse services** – As outlined above, we currently do not have sufficient evidence to specify in detail how services should configure every aspect of their content moderation systems and processes. If we based our recommendations on the limited evidence we do have, which would be drawn from the practices of larger, mainly social media services, we could drive smaller services or non-social media services to adopt practices or seek to achieve outcomes that are not appropriate for their services. This may also result in more onerous expectations on smaller services which may not have the resources or need to match solutions of larger services, potentially undermining their ability to operate, with implications for competition and innovation more widely.
- **Unintended consequences** – Specifying in detail how services should configure all aspects of their content moderation systems and processes may have unintended consequences. For example, telling services to configure their systems and processes in a certain way may not be appropriate for certain services and could decrease user safety, instead of improving it, or could be disproportionate for certain services which could have impacts on competition. By contrast, a more flexible approach, such as in approaches 2 and 3, may allow services to meet the requirements in ways that work for them.

## Approach 2: Specify in detail the outcomes content moderation systems and processes should achieve but leave the design to services

12.32 This section examines whether Codes should specify in detail the outcomes content moderation systems and processes should achieve (i.e. setting detailed KPIs), but leave the design of systems and processes to services. This differs from the first approach as it is less concerned with how a service achieves an outcome and is instead concerned with what it achieves.

12.33 The benefits to this approach would be:

- **Raise standards** – Being specific about the outcomes could potentially raise the standard of content moderation, as it is pushing services to achieve an outcome we deem appropriate for increasing user safety. However, as noted below, there is a general lack of evidence and little consensus on the most effective outcomes to focus on.
- **Ease of compliance** – This approach allows more flexibility than approach 1, as it not concerned with how services achieve an outcome, simply that they achieve it. This approach may be particularly attractive where services are better placed to identify the best approach to achieve desired outcomes and could leave room for services to innovate.

- **Certainty** - Being specific about the outcomes could make it easier for a service to determine whether or not it had complied, reducing its regulatory compliance costs.

12.34 The drawbacks to this approach would be:

- **Diversity of services** – As outlined in approach 1, there is no ‘one-size-fits-all’ approach to content moderation. While this approach may allow some flexibility in how a service achieves an outcome, they would still be required to achieve the specific outcomes set out in Codes. As discussed previously, what is appropriate likely varies from service to service.
- **Lack of evidence** – At this stage, there is a lack of evidence and little consensus on the specific outcomes content moderation systems and processes should be achieving, although we consider exceptions to this general position in Chapter 14.
- **Impacts on smaller and diverse services** – There are similar proportionality concerns to those we considered in relation to approach 1, when it comes to specifying in detail the outcomes content moderation systems and processes should achieve.
- **Unintended consequences** – There is a risk of unintended consequences when setting specific outcomes. It would be difficult to specify outcomes that accurately captured all important dimensions of content moderation systems and processes relevant to removing illegal content and if the specified outcomes did not accurately capture this, they could steer services to behave in a way that was not the most effective.

### Approach 3: require services 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

12.35 This section examines whether Codes should require services to operate a content moderation system and (where relevant) set out the factors to which services should have regard when designing their content moderation systems and processes.

12.36 The benefits to this approach would be:

- **Raise standards** – In comparison to a more prescriptive approach, a broader, more flexible approach could encourage services to achieve compliance in innovative ways, which could raise standards. Nevertheless, there are some drawbacks to this, as highlighted under ‘scope’ below.
- **Diversity of services** – As there is no ‘one-size-fits-all’ approach to content moderation, this approach would allow flexibility in how services choose to meet the measures, allowing them to comply in a way that works for their individual circumstances.
- **Flexibility** - As this is less prescriptive than both approaches 1 and 2 (outlined above), it would leave services with more flexibility to meet the measures in way that is more proportionate and cost-effective for them. In some cases, services may even already be compliant.

12.37 The drawbacks to this approach would be:

- **Scope** – Services might not place appropriate weight on the factors they consider in designing their content moderation systems and processes.
- **Clarity** – there is a risk that this approach would not provide services with sufficient regulatory certainty.

## Choosing an approach

12.38 Each of the approaches set out above has advantages and disadvantages. On balance, our provisional view is that the third approach would be generally more appropriate, although as set out in Chapter 14, we have identified some areas in which we consider we can and should be more prescriptive.

12.39 We therefore propose that our general approach will be to leave services with flexibility as to how to design their content moderation systems (rather than being prescriptive) but set out the factors they should have regard to when considering how to design their systems. We think that it is right to use this approach in most respects at this stage because:

- we do not currently have enough evidence to specify in detail every aspect of how services should configure their content moderation systems and processes (approach 1) or specify in detail the outcomes content moderation systems and processes should achieve (approach 2). In addition, as stated earlier in this chapter, there is no consensus approach to content moderation and different approaches may be more appropriate in different circumstances. Notwithstanding the risks we have identified with approach 3,, we provisionally consider it to be appropriate at this stage. Taking a prescriptive and specific approach at this stage would give rise to a substantial risk of regulatory failure and unforeseen consequences. It could lead to significant disruption in the sector and potentially to increased, rather than decreased, harm to users; and
- our preferred approach allows services greater flexibility on how to behave to achieve compliance and allows services to comply with the measures in ways that may be more proportionate and cost effective for them, while also still, where relevant, setting out the important factors that services should take into account. This is particularly beneficial in this context given the diverse range of services in scope of the regulations and the fast-moving pace of technological development.

## Relationship between Terms of Service and illegal content judgments in content moderation

12.40 We recognise that many service providers will have designed their terms of service and community guidelines to comply with existing laws in multiple jurisdictions and their own commercial needs. For example, if a service has already decided that it wishes to remove all sexual content, we do not think that compliance with the takedown duty creates a need for that service to go on to make a potentially more complex judgement about whether the content concerned amounts to intimate image abuse. The service could simply apply its terms of service.

12.41 The Act allows for service providers to have different terms of service for UK users when compared to users elsewhere in the world. In practice, where the Act requires content to be taken down, this means taken down for UK users.

12.42 To accommodate these principles, we think service providers should have a choice: they may either 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 set 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.

- 12.43 Services may become aware of suspected 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 16. In the same chapter, we also consider whether to propose a means for entities with appropriate expertise and information ('trusted flaggers') to report suspected illegal content to services. In Chapter 14, we identify the automated content moderation (ACM) technology we propose to recommend with a view to identifying further illegal content or suspected illegal content. Services may choose to use other kinds of technology or human content moderators in order to identify suspected illegal content as defined in the Act.
- 12.44 Having set out the general approach we propose to take to provisions regarding content moderation, we now move on to consider the measures we will include in Codes.

## Content moderation systems

---

### Options

- 12.45 We have considered the case for recommending six measures relating to content moderation in our Codes of Practice:
- a) **Measure 1:** All services should have in place content moderation systems or processes designed to swiftly take down illegal content;
  - b) **Measure 2:** Services which are large or multi-risk should set internal content policies having regard to at least the findings of their risk assessment and any evidence of emerging harms on their service;
  - c) **Measure 3:** Services which are large or multi-risk should set performance targets for their content moderation functions and measure whether they are achieving these. These should include targets for both how quickly illegal content is removed and for the accuracy of content moderation decisions. When setting performance targets services should balance the desirability of taking illegal content down swiftly against the desirability of making accurate moderation decisions.
  - d) **Measure 4:** Services which are large or multi-risk should have and apply policies on prioritising content for review, having regard to at least the following factors: virality of content, potential severity of content, the likelihood that content is illegal, including whether it has been flagged by a trusted flagger.
  - e) **Measure 5:** Services which are large or multi-risk should resource their content moderation functions so as to give effect to their internal content policies and performance targets, having regard to at least: the propensity for external events to lead to a significant increase in demand for content moderation on the service; and the particular needs of its United Kingdom user base as identified in its risk assessment, in relation to languages.
  - f) **Measure 6:** Large or multi-risk services should ensure their content moderation teams are appropriately trained.
- 12.46 Below we consider the case for each of these measures in turn.

### Measure 1: Having in place content moderation systems or processes designed to swiftly take down illegal content

---

- 12.47 Services must have in place systems and processes to moderate illegal content in a way that satisfies the requirements contained in the safety duties. To secure this outcome, service providers must ensure these systems or processes are designed such that they remove illegal content swiftly where they become aware of its presence on the service. We have therefore considered including a measure in Codes stipulating that services should have systems or processes designed to swiftly take down illegal content of which they are aware. For this purpose, when a service has reason to suspect that content may be illegal content, it should either:
- make an illegal content judgement in relation to the content and, if it determines that the content is illegal content, swiftly take the content down; or

- where the provider is satisfied that its terms and conditions for the service prohibit the types of illegal content defined in the Act which it has 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.
- 12.48 The option we are considering in this section closely reflects the duty in section 10(3)(b) of the Act (the ‘takedown duty’), which applies in respect of all regulated user-to-user services, rather than to one or other particular type of them. Services must 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*”. However, the requirements remain inherently scalable depending on the service in question and the circumstances.
- 12.49 The systems or processes that the service chooses to put in place must be designed in a way that ensures illegal content of which the service is aware is removed swiftly. What counts as “swift” removal of illegal content in a given case will depend on the circumstances.
- 12.50 The design of this option is not prescriptive as to whether services use wholly or mainly human or automated content moderation processes. Note, however, that in Chapter 14, we are proposing to make some more prescriptive recommendations for certain service providers.
- 12.51 In sum, for the purposes of complying with the takedown duty in the Act, all service providers would need to ensure their content moderation functions are designed to either:
- a) make an illegal content judgement in relation to suspected illegal content and, if it determines that content is illegal content, take the content down swiftly; or
  - b) where a service is satisfied that its terms of service prohibit the types of illegal content defined in the Act which it has reason to suspect exist, consider whether the content is in breach of those terms of service and, if it is, take the content down swiftly.
- 12.52 Accordingly, this option would involve including wording to this effect in our draft Codes. For the avoidance of doubt, a content moderation system may adopt an approach that combines the two processes described above.
- 12.53 Note that in Chapter 14 we make specific proposals which correspond to this proposed measure for content detected using the ACM technology we propose to recommend. The option that we consider here does not affect those proposed measures.

## Costs and risks

- 12.54 The costs of this option will vary by service. For small services with low risks which have few complaints, the costs could be low. Such services may have a process to assess all complaints of illegal content as they arise and take down any illegal content, with the costs being low because they receive few complaints.
- 12.55 For services with significant risks of illegal content, the costs could be considerable as the volume of complaints could be high and the content moderation systems and processes may need to be substantial.
- 12.56 However, we consider that the option outlined here is really the minimum that services would need to adopt in order to determine complaints and meet what is a specific requirement in the safety duty, albeit subject to a proportionality threshold, to take down



illegal content of which they become aware. Incurring these costs is therefore necessary to meet the requirements of the Act.

## Rights impact

- 12.57 Content moderation is an area in which the steps taken by services as a consequence of the Act may have a significant impact on the rights of individuals and entities - in particular, to freedom of expression under Article 10 ECHR and to privacy under Article 8 of the European Convention on Human Rights ('ECHR').
- 12.58 Articles 8 and 10 ECHR are 'qualified' rights, interference with which may be justified on specified grounds, including relevantly the prevention of crime, the protection of health and morals, and the protection of the rights of others.
- 12.59 In considering whether impacts on these rights are proportionate, our starting point is to recognise that Parliament has determined that services should take proportionate steps to protect users from illegal content. They must, in particular, take proportionate steps to secure that such content is taken down when the service becomes aware of it. Parliament has also identified a series of offences as 'priority offences' in this context.
- 12.60 We therefore take it that a substantial public interest exists in measures which aim to reduce the prevalence and dissemination online of priority illegal content. That public interest relates to each of the prevention of crime, the protection of health and morals, and the protection of the rights of others.
- 12.61 The detection and removal of illegal content acts directly to prevent crime in a number of ways, such as by deterring users from posting such illegal content. It may prevent other users from accessing it (and so potentially committing further offences either because accessing the content is itself an offence, or because the content encourages or assists in the commission of offences). It similarly acts to protect public morals, including by preventing users inadvertently encountering illegal content online.
- 12.62 The removal of some kinds of illegal content also acts directly to protect the rights of victims, for example those depicted in 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' rights under Article 8 ECHR and protects victims' personal data.
- 12.63 We consider the potential impacts on users' freedom of expression and privacy in this light.

## Freedom of expression

- 12.64 An interference with the right to freedom of expression must be prescribed by 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. Potential interference with users' freedom of expression arises where content is taken down because the service considers it to be illegal content, particularly if that judgement is incorrect. As set out above, however, our starting point is that Parliament has determined that services should take proportionate steps to protect UK users from illegal content. Of course there is some risk of error in them doing this, but that risk is inherent in the scheme of the Act.



- 12.65 Services have incentives to limit the amount of content that is wrongly taken down, to meet their users' expectations and to avoid the costs of dealing with appeals.
- 12.66 In addition, there could be a risk of a more general 'chilling effect' if users were to avoid use of services which have implemented a more effective content moderation process as a result of this option. However, we do not consider that any such effect would be significant, given that many UK users already use services which have implemented content moderation processes.
- 12.67 A greater interference would arise if the service, because of the Act, chose to adopt terms of service which defined the content it prohibited more widely than is necessary to comply with 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 prohibit content that is not or might not be illegal content, so long as they abide by the Act. Nothing in this option asks that services take steps against any content other than illegal content. Services have incentives to meet their users' expectations in this regard, too.
- 12.68 The duty for services to treat illegal content appropriately is a function of the Act, and not of this measure. This option is designed in a way that is not prescriptive about how illegal content is to be moderated, just that the provider's systems or processes are designed such that they remove illegal content swiftly where they become aware of its presence on the service. It does not involve services taking any particular steps in relation to content of which they are not aware.
- 12.69 Impacts on freedom of expression could in principle arise in relation to the most highly protected forms of content, such as religious or political expression, and in relation to kinds of content that the Act seeks to protect, such as content of democratic importance and journalistic content. However, we consider there is unlikely to be a systematic effect on these kinds of content.
- 12.70 Where a service takes down content on the basis that it is illegal content, complaints procedures operated pursuant to section 21(2) of the Act allowing for the user to complain and for appropriate action to be taken in response may also mitigate the impact on their rights to freedom of expression.<sup>50</sup>

## Privacy

- 12.71 An interference with the right to privacy must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest. Again, in order to be 'necessary', the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.
- 12.72 Insofar as services use automated processing in content moderation, we consider that any interference with users' rights to privacy under Article 8 ECHR would be slight. Such processing would need to be undertaken in compliance with relevant data protection legislation (including, so far as the UK GDPR applies, rules about processing by third parties or international data transfers).
- 12.73 Review of suspected illegal content by human moderators, including those employed by a contracted third party, involves more significant potential impacts on privacy both of the user and persons mentioned or depicted in the content. However, that review (and the

---

<sup>50</sup> See Chapter 16 (Reporting and complaints).

associated interference) is for the purpose of ensuring illegal content is taken down accurately for the purpose of the safety duty.

- 12.74 Interference with users' or other individuals' privacy rights may also arise insofar as the proposed measure would lead to reporting to reporting bodies or other organisations in relation to illegal content. In particular, section 67 of the Act makes provision which (when brought into force) will require providers of regulated U2U services to report detected and unreported CSEA content to the Designated Reporting Body housed in the NCA (as further specified in the Act and to be specified in regulations made by the Secretary of State under section 68 of the Act). Providers may also have obligations to report CSEA content in other jurisdictions, or may have voluntary arrangements in place. For example, US providers are obliged to report to NCMEC under US law when they become aware of child sexual abuse on their services.
- 12.75 In part, any such interference results from the duties created by the Act or by existing legislation in other jurisdictions. In particular, where users or other individuals are correctly reported pursuant to the Act because they are suspected of committing the offence related to the CSEA content, any interference with their rights 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.
- 12.76 However, we have considered the extent to which the inclusion of this measure in our Codes of Practice as a recommended measure for the purpose of complying with providers' illegal content safety duties might give rise to additional interference.
- 12.77 Errors in content moderation decisions, whether made by automated technology or by humans, could result, in effect, in individuals being incorrectly reported to reporting bodies or other organisations, which would represent a potentially significant intrusion into their privacy. It is not possible to assess in detail the potential impact of incorrect reporting of users: the number of users affected would depend on what systems and processes the service implemented.
- 12.78 However, we do not consider it proportionate to expect all services, including very low risk, small and micro-businesses, to build in extra systems and processes to avoid accidental incorrect reporting to reporting bodies. (We consider what more might be needed for larger and riskier services below.) Reporting bodies have processes in place to triage and assess all reports received, ensuring that no action is taken in cases relating to obvious false positives. These processes are currently in place at NCMEC and will also be in place at the Designated Reporting Body in the NCA, to ensure that investigatory action is only taken in appropriate circumstances.

## **Provisional conclusion**

- 12.79 All services must have proportionate systems and processes designed to take down illegal content swiftly. Our proposed approach is to recommend that all services operate these systems and processes, but without specifying how content is removed.
- 12.80 We are provisionally recommending that all regulated U2U services should have systems or processes designed to take down illegal content of which they are aware swiftly. For this purpose, when a service has reason to suspect that content may be illegal content, it should either:

- a) make an illegal content judgement in relation to the content and, if it determines that the content is illegal content, take the content down swiftly; or
- b) where the provider is satisfied that its terms and conditions for the service prohibit the types of illegal content defined in the Act which it has reason to suspect exist, consider whether the content is in breach of those terms of service and, if it is, take the content down swiftly.

12.81 The costs of this measure will vary by service. Regardless of the level of the costs for a particular service, we consider this measure proportionate. This is because we see having content moderation systems or processes in place that are designed to take down illegal content swiftly as being necessary to meet the requirements of section 10(3)(b) of the Act. If it is necessary to meet those requirements, it must be a proportionate way to meet the requirements.

12.82 Overall, we consider that the impacts of this proposed measure on users' rights to freedom of expression under Article 10 ECHR, and to privacy under Article 8 ECHR are justified by the substantial public interest in the prevention of crime, the protection of health and morals, and the protection of the rights of victims and children that this proposed measure is designed to achieve, and are proportionate to the anticipated benefits of the measure from reducing the prevalence and dissemination of illegal content. We also do not consider that there is a less intrusive way of achieving these aims.

12.83 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

## **Measure 2: Services which are large or multi-risk should set internal content policies having regard to at least the findings of their risk assessment and any evidence of emerging harms on their service**

---

### **Effectiveness**

12.84 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's rules about what content is allowed and what is not. These are often in the form of terms of service and/or community guidelines. Internal content policies are usually more detailed versions of external content policies which 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. Once internal content policies are set, they can be used as a guide for enforcement by content moderators and other relevant teams, as well as designers of automated systems to assist in identifying potential content breaches.<sup>51</sup>

---

<sup>51</sup> Alan Turing Institute, 2021. [Understanding online hate: VSP Regulation and the broader context](#); Meta, 2021. [What's Allowed on Our Platforms? Find Out in Episode 2 of Video Series, Let Me Explain](#). [accessed 3 August 2023]; Twitch, 2022. [Transparency Report](#). [accessed 3 August 2023]; Trust and Safety Professional Association, no date. [Policy Development](#). [accessed 3 August 2023]; Khoury College at Northeastern University, no date. [Content Moderation Techniques](#). [accessed 3 August 2023]; Twitter, no date. [Our approach to policy development and enforcement philosophy](#). [accessed 3 August 2023]; Bumble, no date.

- 12.85 Evidence from industry stakeholders suggests that there is a broad consensus that setting internal content policies is a necessary first step to establishing an effective content moderation system for some services. For example, several large and medium platforms publicly state that content policies play a key part in keeping users safe online.<sup>52</sup> There is a strong in-principle argument that where services are larger or higher risk and therefore need to moderate large volumes of diverse content, it is important that they have clear content moderation policies in order to ensure consistency, accuracy and timeliness of decision making.
- 12.86 This suggests that for large or risky services the existence of internal content policies is a pre-condition for moderating content effectively. Given the evidence we have presented above on the role content moderation plays in reducing harm, we therefore consider that a measure recommending large services or services that face significant risks set internal content policies would result in material benefits.
- 12.87 We also consider that there would be significant benefits in recommending that services have regard to at least their risk assessments and evidence of emerging harms when setting their policies. Both of these data sources would provide evidence about the challenges services' content moderation functions face. It is reasonable to infer that such data would enable services to make higher quality decisions about what to put in their internal content moderation policies. This should improve the quality of these policies and by extension improve the performance of services' content moderation systems, thereby reducing harm to users.

## Costs and risks

- 12.88 Services that do not currently have internal content policies would incur the costs of developing them. This could take a small number of weeks of full-time work and involve legal, regulatory, as well as different ICT staff, and online safety/ harms experts. In some cases, services may use external experts which could increase costs. Agreeing new policies may also take up senior management's time which would add to the upfront costs. For most services we expect these costs to be in the thousands of pounds, although larger/riskier services may require more complex content policies which may increase costs. In addition there may be some small ongoing costs to ensure these policies remain up to date over time.

## Rights impact

### Freedom of expression

- 12.89 The reasoning on the right to freedom of expression set out in relation to Measure 1 above applies equally to this option.

---

[Guidelines](#). [accessed 3 August 2023]; Discord, no date. [Discord Community Guidelines](#). [accessed 3 August 2023]; Niantic, no date. [Niantic Player Guidelines](#). [accessed 3 August 2023]; Roblox, no date. [Safety & Civility at Roblox](#). [accessed 3 August 2023]; TikTok, no date. [Community Guidelines](#). [accessed 3 August 2023]; YouTube, no date. [How does YouTube manage harmful content?](#) [accessed 3 August 2023].

<sup>52</sup> TikTok, 2019. [Creating Policies for Tomorrow's Content Platforms](#). [accessed 3 August 2023]; YouTube, 2019. [The Four Rs of Responsibility, Part 1: Removing harmful content](#). [accessed 3 August 2023]; Meta, 2020. [Facebook's response to Australian Government consultation on a new Online Safety Act](#). [accessed 3 August 2023]; Mid-Sized Platform Group, 2022. [Mid-Sized Platform Group – Online Safety Bill Recommendations](#). [accessed 3 August 2023]; Twitter, no date. [The Twitter Rules](#). [accessed 3 August 2023].

- 12.90 This option is designed in a way that does not tell services how to moderate illegal content, just that there are internal content policies outlining how to moderate it.
- 12.91 There is some risk that in writing their policies, services which align their terms and conditions with the definition of illegal content in the Act may over-generalise in a way which leads to over moderation. However, we consider that this risk arises equally if we were to not propose this option, since content moderators operating without any internal guidance may also over-generalise or be overly cautious.
- 12.92 Where services are likely to be dealing with large volumes of content, the process of considering these matters in advance and preparing a policy would tend to improve internal scrutiny, and improve the consistency and predictability of decisions, in a way which we think would also tend to protect users' rights to freedom of expression.

## Privacy

- 12.93 To the extent that, in setting content policies, services describe or define the content they are prohibiting in a way which involves reference to information in respect of which a user would have a reasonable expectation of privacy, or to personal data, users' rights in relation to these would be engaged.
- 12.94 However, that review (and the associated interference) is for the purpose of ensuring illegal content is taken down accurately for the purpose of the safety duty.
- 12.95 Where services are likely to be dealing with large volumes of content, the process of considering these matters in advance and preparing a policy would tend to improve internal scrutiny, and improve the consistency and predictability of decisions, in a way which we think would also tend to protect users' privacy and personal information rights.

## Provisional conclusion

- 12.96 Multi-risk services pose significant risks to their users. We consider that the benefits of applying this measure to them are therefore likely to be material. Our analysis suggests that for services that face significant risks, the presence of internal content policies is an important part of an effective content moderation systems which reduces harm to users. Such services are unlikely to be able to moderate content effectively without such policies. As we have explained, the absence of effective content moderation materially increases the risks of illegal content being disseminated on services. At the same time, the costs of this measure are likely to be relatively small for many multi-risk services. We therefore consider that it would be proportionate to apply the measure to all multi-risk services.
- 12.97 The benefits of recommending this proposed measure to large services with low risks of illegal harm would not be as great, as there would be less scope to reduce harms from illegal content. However, we still consider that having internal content moderation policies in place for such services will still have important benefits for users. This is partly because such services have the potential to affect a lot of users, and also because the nature of illegal content can change over time meaning that even if a large U2U service is low risk currently, this could change in the future. We anyway anticipate that large services will generally identify themselves as multi-risk, as their large reach tends to increase the impact of any illegal content. We also note that many large services are likely to have content policies in place already and, if they do not, are likely to have sufficient resources to develop them. We therefore consider that it would be proportionate to apply this measure to large services with low risks.

- 12.98 We are not proposing to recommend this measure for smaller and lower risk services. We consider the benefits of an internal content moderation policies are likely to be materially smaller for services which are neither large nor face material risks. They are unlikely to face large volumes of content they need to assess. So even though the costs of this measure are low, we do not propose to recommend it for such services.
- 12.99 In light of the analysis above we propose that our Codes should recommend that large services and multi-risk services should set internal content moderation policies having regard to at least the findings of their risk assessment and any evidence of emerging harms on their platform.
- 12.100 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

### Measure 3: Services which are large or multi-risk should set performance targets for their content moderation functions

---

#### Effectiveness

- 12.101 We have considered the case for recommending the following measure:
- a) Services which are large or multi-risk should set performance targets for their content moderation functions and track whether they are meeting these. These should include targets for both how quickly illegal content is removed and for the accuracy of content moderation decisions. When setting targets services should balance the need to take illegal content down swiftly against the need to make accurate moderation decisions. They should measure their performance against their targets.
- 12.102 We understand that many services set performance targets for the operation of their content moderation functions and measure whether they are achieving these. For example, in response to the 2022 Illegal Harms Call for Evidence, OnlyFans told us that, within two minutes 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.<sup>53</sup> [CONFIDENTIAL X].<sup>54</sup>
- 12.103 We consider that setting performance targets and measuring whether they are achieving these is likely to deliver important benefits. Where services are clear about the content moderation outcomes they are trying to achieve and measure whether they are achieving them, it stands to reason that they will be better able to plan how to configure their systems to meet these goals and better able to optimise the operation of these systems.
- 12.104 The importance of measuring performance against targets is reinforced by the evidence we have collected from stakeholders. A number of stakeholders, in response to the 2022 Illegal Harms Call for Evidence, stressed the importance of reviewing the performance of content moderation systems, including BSR, Global Partners Digital and the Ombudsman Services Internet Commission, with the latter noting that, "Moderators and automated processes can

---

<sup>53</sup> OnlyFans, 2022. [OnlyFans response to the 2022 Illegal Harms Call for Evidence](#).

<sup>54</sup> [CONFIDENTIAL X].



remove too much or too little content. Holding regular quality assurance sessions where a sample of decisions can be checked, and feedback could be provided particularly on contentious issues should be part of a running dialogue in the organisation." This is also reflected by some other civil society organisations, including the Santa Clara Principles and the Trust & Safety Professional Association.

- 12.105 Consistent with the general approach described earlier in this chapter, we do not propose to stipulate what the performance targets services should set. However, under the option we are looking at we *would* propose that at a minimum these should include targets relating to the time within which services review or remove illegal content and targets relating to the accuracy of content moderation decisions.
- 12.106 Some services record a wide range of metrics in relation to content moderation systems and processes. While many services record the same or similar metrics, there is considerable variation in precise definitions and naming conventions. The Trust & Safety Professional Association (TSPA) draws together these various metrics into five broad categories: enforcement volume metrics,<sup>55</sup> time-based metrics;<sup>56</sup> quality metrics,<sup>57</sup> appeals metrics,<sup>58</sup> and other metrics.<sup>59</sup>
- 12.107 We consider that there would be important benefits to services setting both time based and quality/accuracy based targets for their content moderation teams and having regard to the desirability of striking a balance between timeliness and accuracy of decision making when setting their performance targets. Users are only protected if decisions are made in a timely way. Therefore there is a clear benefit to services having regard to the need for timely review of potentially harmful content when setting their performance targets. At the same time, accuracy of decision making is also important and there is a strong case that a focus on speed of decision making should be balanced with a focus on accuracy. A disproportionate focus on speed of content removal could lead to pressure on systems which results in poorer quality decisions, which in turn could lead to a decrease in accuracy. This is an issue that has already been levelled against some services. 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

---

<sup>55</sup> 'Enforcement Volume Metrics' represent counting events that are part of the moderation process, such as capturing the volume of content flagged for review, the volume of content closed by a service's content moderation system, and the number of instances where a moderation action was taken. Trust & Safety Professional Association, no date. [Metrics for Content Moderation](#). [accessed 3 August 2023].

<sup>56</sup> 'Time Based Metrics' are based on the amount of time taken to perform various parts of the content moderation process, such as review time, response time, removal time and time to action, i.e. the time between content being uploaded or created and a completed decision about whether the content is violating. Trust & Safety Professional Association, no date. [Metrics for Content Moderation](#). [accessed 3 August 2023].

<sup>57</sup> 'Quality Metrics' are generally based on re-checks of previous reviews by either the existing review teams, subject matter experts, or dedicated quality reviewers. Trust & Safety Professional Association, no date. [Metrics for Content Moderation](#). [accessed 3 August 2023].

<sup>58</sup> 'Appeals Metrics' involve re-checks of previous reviews by either the existing review teams, subject matter experts, or dedicated quality reviewers based on appeals, such as overturns and overturn rate, successful appeal rate, and time to resolution. Trust & Safety Professional Association, no date. [Metrics for Content Moderation](#). [accessed 3 August 2023].

<sup>59</sup> 'Other Metrics' tend to be less directly tied to day-to-day operational decisions, such as prevalence, cost and impressions. Trust & Safety Professional Association, no date. [Metrics for Content Moderation](#). [accessed 3 August 2023].

decisions". Google noted similar concerns in relation to NetzDG and also in its response to the Australian Government's Consultation on Online Safety Reforms.

## Costs and risks

- 12.108 Services 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 determine how long and how many views there have been of content that is subsequently found to be illegal, or for tracking the time between when content is reported and when it is assessed or taken down if found to be violative. There would also be ongoing costs. This would include data storage costs. More significantly, to assess the accuracy of content moderation decisions, services are likely to need to take a sample of those decisions and re-assessing them. There could therefore also be significant on-going costs from this measure.
- 12.109 We are not able to quantify these costs with any precision. They would depend in part on the complexity of the targets services set and the volume of content that was assessed.
- 12.110 There is a risk that setting performance targets could give rise to perverse incentives. For example, in principle there is a risk that unduly rigid targets could cause services to make sub-optimal decisions about which pieces of content to prioritise for review. However, we consider that our proposal is structured in such a way as to substantially mitigate this risk, given that we are allowing services flexibility for how to structure their targets and have explicitly set out that services should balance speed and accuracy of decision making.

## Rights impact

- 12.111 Our assessment of the rights impacts associated with Measure 1 also applies to this option in that moderating content can infringe users' rights to both free expression and privacy. The risks to both can be increased by the addition of performance targets. A performance target relating to speed can cause moderators to try to take decisions quickly, increasing the risk of error and impacts on freedom of expression. A performance target relating to accuracy could, in some cases, incentivise moderators to seek to review more content than they need to, to be more sure that decisions are correct.
- 12.112 However, this option is designed to cause services to balance the need to take illegal content down swiftly with the need to make accurate moderation decisions. In particular, it does not specify a time within which decisions must be made, so the option should not put pressure on moderators to act so fast as to put users' rights to freedom of expression at risk.
- 12.113 The risks to privacy set out in relation to Measure 1, arising from the possibility that services may report detected illegal content to reporting authorities, are particularly acute where services are likely to be moderating content in large volumes. Whether automated technology is used, turnover of moderation staff, time pressures, seniority and experience of the person concerned can all affect the likelihood of error. We consider that the setting and monitoring of accuracy targets as a part of this option, also acts as a safeguard for users' rights to freedom of expression.

## Provisional conclusion

- 12.114 For services that have material volumes of content to assess, we consider there would be important benefits from setting performance targets for their content moderation functions and tracking whether they are met. As we explain above, we consider that services that follow this measure are more likely to operate effective content moderation systems. As we



have shown, the evidence suggests that effective content moderation plays a hugely important role in mitigating the risk of harm to users meaning the measure would have important benefits. As with measure 2, these benefits will be greatest for services that are either large or multi-risk.

- 12.115 The costs of this measure are somewhat unclear. However, on balance, we consider that even in the context of this uncertainty, the benefits are likely to be sufficiently important to justify this proposal for large services and multi-risk services given the fundamental role effective content moderation plays in protecting users from harm. That this proposed measure is proportionate is also consistent with Ofcom not proposing to be prescriptive on the details of the performance targets set or how they are achieved. This leaves scope for services to tailor these targets according to the risks they identify and the specific operation of their services. This flexibility helps ensure that services can design performance targets and systems that are proportionate. Moreover, the measure is in line with common practice in industry and any concerns about cost are mitigated by the fact that we are only targeting it at large services and multi-risk services.
- 12.116 We are not proposing to recommend this measure for smaller and lower risk services, because it is less clear the benefits are great enough given the lower volume of content such services need to assess.
- 12.117 We therefore propose that our Codes should recommend that:
- a) Services which are large or multi-risk should set performance targets for their content moderation functions and track whether they are meeting these. These should include targets for both how quickly illegal content is removed and for the accuracy of content moderation decisions. When setting targets services should balance the desirability of taking illegal content down swiftly against the desirability of making accurate moderation decisions. They should measure their performance against their targets.
- 12.118 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

## Measure 4: Services which are large or multi-risk should have and apply policies on prioritising content for review

---

- 12.119 Below we set out our analysis of the case for recommending the following measure in codes:
- a) **Measure 4:** Large or multi-risk services should have and apply policies on prioritising content for review. In setting the policy, the provider should have regard to at least the following factors: virality of content, potential severity of content, the likelihood that content is illegal, including whether it has been flagged by a trusted flagger.

### Effectiveness

- 12.120 Given the immense amount of content posted on them, large U2U services often get huge volumes of content flagged to them as potentially illegal or otherwise harmful. Smaller multi-risk services, too, are likely to have many different types of content to moderate at once. This means both types of service face difficult decisions about what content to prioritise for review. The decisions they take about what content to prioritise can have a material impact on the amount of harm a piece of illegal content does to people. For

example, if a service chooses 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 being viewed by large numbers of people, this decision would result in significant harm to users.

- 12.121 Many services use systems and processes to help them prioritise content for review. Services dealing with content moderation on a large scale do not typically review content in chronological order but consider a range of factors, including: the virality of the content, its severity, and the circumstances surrounding it becoming known to the platform (for example, whether or not as a consequence of a user report or other complaint).<sup>60</sup>
- 12.122 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.<sup>61</sup> Additionally, Facebook prioritises items based on how confident an algorithm is that moderators will agree that the content is violative and also on the ‘severity’ or ‘egregiousness’ of a suspected violation – arguably linked to the degree of harmfulness. However, one side effect of this is that relatively less popular or less harmful items may remain available for long periods of time.<sup>62</sup>
- 12.123 Prioritising content also relies on services making trade-offs between a number of important goals, including harm reduction, users’ freedom of expression, and user experience. We currently think services are usually best placed to make these decisions based on their individual needs, although in Chapter 14 we set out some specific content detection processes which we consider ought to be established.
- 12.124 We consider that where a service adopts a prioritisation framework which considers the factors listed above (as well as other factors they identify as relevant) this is likely to result in high quality decisions about what content to prioritise for review. Logically, we would expect this to result in a material reduction in harm to users compared to a counterfactual in which services simply reviewed complaints in a chronological order, thereby delivering significant benefits. The benefits of having such a framework would likely be materially smaller for services which are neither large nor face material risks. This is because they are likely to receive materially fewer complaints for review.
- 12.125 We explain below why each of the prioritisation criteria covered by our option are important and relevant:

## Virality of content

- 12.126 Virality is a term used to describe the degree to which online content spreads easily and/or quickly across many online users, alongside how much engagement and/or views a piece of content receives (i.e. ‘shares’, ‘likes’, views’, etc.).
- 12.127 If illegal content is going viral, i.e. reaching a higher number of users than is typical within a given timeframe, it has the potential to cause harm to larger audiences. The purpose of the

---

<sup>60</sup> Cambridge Consultants, 2019. [Use of AI in Content Moderation](#). [accessed 3 August 2023]; Meta, 2020. [How We Review Content](#). [accessed 3 August 2023]; Google, 2020. [Information quality and content moderation](#). [accessed 3 August 2023]; Meta, 2022. [How Meta Prioritises Content for Review](#). [accessed 3 August 2023]; [CONFIDENTIAL X].

<sup>61</sup> Ofcom, 2023. [Content moderation in user-to-user online services: An overview of processes and challenges](#), p.7. [accessed 25 September 2023].

<sup>62</sup> Ofcom, 2023. [Content moderation in user-to-user online services: An overview of processes and challenges](#), p.20. [accessed 25 September 2023].

Act is to make the use of regulated internet services safer for individuals in the United Kingdom.<sup>63</sup> We therefore provisionally think services will achieve better outcomes for users if they have regard to virality when prioritising content.

12.128 We know that several of the larger services consider ‘virality’ of content when prioritising content for review, including both the ‘likely’ virality and ‘actual’ virality.<sup>64</sup>

12.129 However, we note that it is important to balance virality alongside other factors, including those listed here, as prioritising virality alone may mean other harms are missed. For example, CSAM does not typically go viral but is high-severity content. Similarly, content constituting harassment and threats or intimate image abuse may be targeted at an individual and may not go viral, but can be high-severity for the individual concerned – this can be particularly harmful to women and girls.

12.130 It should also be noted that some services in the 2022 Illegal Harms Call for Evidence, such as OnlyFans, told us they design their platforms so that content cannot go viral.<sup>65</sup> Nevertheless, services may still need to consider how quickly content is spreading or how many views/how much engagement a piece of content is receiving.

### Severity of content, including whether it is likely to relate to a priority illegal harm

12.131 We know that several services already consider the severity (or egregiousness) of harm when prioritising content for review.<sup>66</sup> Some harms may be considered to have higher severity than others, such as those that have a degree of immediate direct harm compared to those that do not. For example, the immediacy of livestreamed illegal content, such as terrorist attacks, may require real time moderation, or moderation that is faster than non-livestreamed content, so it may be appropriate to prioritise these.<sup>67</sup> All else being equal, services reviewing and removing high severity illegal content promptly is likely to reduce harm to people in the UK.

12.132 However, even within certain harms, there may be degrees of severity that need to be considered. For example, in its report into online hate, the Alan Turing Institute noted that “different types of online hate inflict different degrees and types of harm”. With this in mind, it might be that services focus on the potential severity or impact of harm to help them prioritise content for review, so when they come to review it they can carefully consider other concerns, such as context, freedom of expression, etc.<sup>68</sup>

12.133 In response to the 2022 Illegal Harms Call for Evidence, Refuge provided examples of children and women who have suffered online abuse waiting months or years for any action to be taken, if it is taken at all. It cited its research which showed that “survivors are

---

<sup>63</sup> Section 1(1) of the Act.

<sup>64</sup> Meta, 2020. [How We Review Content](#). [accessed 3 August 2023]; Ofcom, 2023. [Content moderation in user-to-user online services: An overview of processes and challenges](#).

<sup>65</sup> OnlyFans, 2022. [OnlyFans response to the 2022 Illegal Harms Call for Evidence](#). [CONFIDENTIAL X].

<sup>66</sup> Ofcom, 2023. [Content moderation in user-to-user online services: An overview of processes and challenges](#). [accessed 25 September 2023].

<sup>67</sup> Christchurch Call, no date. [The Christchurch Call to Action](#). [accessed 25 August 2023].

<sup>68</sup> The Alan Turing Institute, 2022. [The Alan Turing Institute response to Illegal Harms Call for Evidence](#). [accessed 25 August 2023].

experiencing tech abuse for extended periods of time. On average, survivor survey respondents endured tech abuse for at least six months.”<sup>69</sup>

- 12.134 ‘Severity’ is also one of the three factors the UK Government used to determine its list of priority illegal offences and services should therefore consider these offences as high-severity.<sup>70</sup> However, services may determine that harms outside the list of priority illegal offences have a high-severity on their platform and should be prioritised in some circumstances – for example, Ofcom will be consulting in due course on content likely to be harmful to children.

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

- 12.135 All else being equal, prioritising content for review where the signals available to the service suggest that there is a high likelihood that it is illegal should increase the speed with which illegal content is removed, thereby reducing harm to users. Reasons to suspect that content is illegal can arise in a number of different ways. Most obviously, users may complain about it. Their reports are likely to be the first and a very valuable way in which services may find out about illegal content, particularly for those services which are not making extensive use of proactive detection methodologies. However, we recognise that users are not always very good at correctly identifying breaches of services’ content policies.<sup>71</sup>
- 12.136 Dedicated reporting channels (DRCs), used by trusted flaggers<sup>72</sup> and Internet Referral Units<sup>73</sup>, are sometimes used by services to flag potentially illegal or violative content for review.
- 12.137 Trusted flaggers can include internal teams, law enforcement, public sector organisations, civil society and private entities, and can offer particular expertise in notifying the presence of potentially illegal content on their website, which may result in higher quality flags or reports and potentially swifter removal of illegal content.<sup>74</sup> In Chapter 16 we consider whether to recommend that services establish a DRC for certain trusted flaggers relating to fraud.
- 12.138 Complaints are already commonly used to help prioritise content for review, and they can potentially flag illegal content that other content moderation functions may have missed. Where services have DRCs in place, the fact that a complaint comes from a trusted flagger or

---

<sup>69</sup> Refuge, 2022. [Refuge response to the 2022 Illegal Harms Call for Evidence](#).

<sup>70</sup> 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 2 August 2023].

<sup>71</sup> 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 26 September 2023]; 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 26 September 2023].

<sup>72</sup> Trusted flaggers are individuals, NGOs, government agencies, and other entities that have demonstrated accuracy and reliability in flagging content that violates a platform’s Terms of Service. As a result, they often receive special flagging tools such as the ability to bulk flag content.

<sup>73</sup> Internet Referral Units are government-established entities responsible for flagging content to internet platforms that violates the platform’s Terms of Service. Examples include the [EU Internet Referral Unit \(EU IRU\)](#) and the UK’s [Counter Terrorism Internet Referral Unit \(CTIRU\)](#).

<sup>74</sup> European Commission, 2017. [Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms](#). [accessed 8 August 2023].

another expert body is of obvious relevance in determining what priority to give it as, all other things being equal, such complaints are likely to be accurate and to reflect the trusted flagger's assessment of harm. They have significant potential to reduce harm to users.

- 12.139 In Chapter 14, we consider certain kinds of automated technology which are associated with a high likelihood that content they identify is illegal. Services may use other kinds of detection, whether human or automated, to identify content as suspected illegal content with varying degrees of certainty. The likelihood that the content is illegal is self-evidently relevant to whether further review is needed and how quickly it should take place.

## Costs and risks

- 12.140 The creation of a prioritisation framework would not in and of itself have an impact on the overall amount of content flagged to services as potentially illegal. However, there would be costs of designing and applying the prioritisation policy. The largely one-off costs of designing the prioritisation policy may take a small number of weeks of full-time work and involve legal, regulatory, as well as different ICT staff, and online safety/ harms experts, and agreeing the policy would likely need input from senior management. Applying that prioritisation policy could require system changes. For example, this might involve ensuring the virality of content 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. There are likely to be some smaller ongoing costs in ensuring that the prioritisation policy is still reflected in system design, and in reviewing it when appropriate. These costs are mitigated by the proposed measure not specifying exactly how services should prioritise content, giving services some flexibility in what they do.
- 12.141 As the amount of content reviewed may not change, it is not clear that establishing a framework for prioritising what content they review having regard to the criteria set out here would impose other material ongoing content moderation costs on services compared to a counterfactual in which they simply reviewed complaints chronologically. Indeed, to the extent that services do not do this already, having a clear prioritisation framework may help them deploy their resources more efficiently.

## Rights impact

- 12.142 Our assessment of the rights impacts associated with having a content moderation function is set out above in relation to Measure 1. We do not consider that setting and applying a prioritisation policy would necessarily have any additional impacts on those rights. To the extent that it meant that harm would be a factor in services' decision making and that more users were better protected against harm, it is likely to result in a more proportionate approach to content moderation by the service, and therefore tend to safeguard users' rights.

## Provisional conclusions

- 12.143 For services that have a large quantity of potentially illegal content to review, there are likely to be significant benefits from prioritising that review in the way we propose, to reduce the harm from illegal content. While there are likely to be one-off costs of establishing a prioritisation system, we have not identified any large ongoing costs associated with the option. We consider that the benefits of adopting a prioritisation framework are sufficiently

important to justify the costs of doing so. This view is reinforced by the fact that our analysis suggests a number of services already use prioritisation frameworks of this sort. This is consistent with the costs being proportionate for those services. As the proposed measure does not specify exactly how services should prioritise content, services have some flexibility to shape their approach to be proportionate to the risk that are on their service.

12.144 We therefore propose to recommend that large or multi-risk services should have and apply policies on prioritising content for review. In setting its policy, a service should have regard to at least the following factors: virality of content, potential severity of content, the likelihood that content is illegal, including whether it has been flagged by a trusted flagger.

12.145 As set out above, the benefits of having a prioritisation framework are likely to be materially smaller for services which are not large and are low risk or single risk. This is because such services are not likely to need to review nearly as much or as diverse potentially illegal content and are therefore less likely to face difficult and consequential prioritisation decisions. At this time, we are therefore not proposing to extend this recommendation to such services.

12.146 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

## **Measure 5: Services which are large or multi-risk should resource their content moderation functions sufficiently**

---

12.147 Given the immense amount of content posted on them, large U2U services often get huge volumes of content flagged to them as potentially illegal or otherwise harmful. Smaller multi-risk services, too, are likely to have many different types of content to moderate at once. This means both types of service are unlikely to be able to keep users safe merely by securing that, for example, whichever member of senior management is available reviews complaints when they come in. They are likely to need dedicated resources of some kind, and are likely to need to adjust the overall resources available, and how they are deployed, depending on what is happening on their service.

12.148 We have considered the case for recommending the following measure in our Codes:

- a) Services which are large or multi-risk should resource their content moderation functions so as to give effect to their internal content policies and performance targets, having regard to at least: the propensity for external events to lead to a significant increase in demand for content moderation on the service; and the particular needs of its United Kingdom user base as identified in its risk assessment, in relation to languages.

12.149 We set out our analysis and findings below.

### **Effectiveness**

12.150 Where content moderations functions are adequately resourced one would expect this to enable them to review potentially illegal content more quickly and make more accurate decisions as to whether to remove it. We therefore consider that where content moderation functions are adequately resourced this will deliver very significant and important benefits.

12.151 This view is reinforced by the fact that, as we explained in Volume 2 chapter 6U, research has shown that the reduction of staff for content moderation in a large service led to a major increase in the quantity of antisemitic content on the service.

- 12.152 Setting objectives in relation to time and accuracy of a U2U moderation function as set out above would not protect users unless the service also set out to resource itself sufficiently, and deploy its resources effectively, so as to meet them. We therefore consider there would be significant benefits to services resourcing its content moderation function so as to meet these performance targets.
- 12.153 We do not at this stage think it would be beneficial for us to specify in detail how services should resource their content moderation functions. However, we do consider that there are factors to which services should have regard when deciding how to resource their content moderation function, and that considering these is likely to result in important benefits.
- 12.154 We explain below the factors we think services should consider and why each factor is important.

### Meeting spikes in demand for content moderation driven by external events

- 12.155 Evidence suggests that for their content moderation function to be effective, services also need to build in flexibility. 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.<sup>75</sup> In response to the 2022 Illegal Harms Call for Evidence, BSR stressed the importance of services “investing in the capability to scale-up/scale-down on short notice to respond to crisis events that can result in sudden spikes in illegal content.”<sup>76</sup> We therefore consider that there would be important benefits if services had regard to the possibility of demand for content moderation surging in response to external events and resourced their content moderation accordingly.
- 12.156 Information obtained from services’ risk assessments, tracking evidence of new kinds of illegal content and other relevant sources of information, could be used to understand where and when such occurrences might happen. In Chapter 8, paragraphs 8.139 to 8.149, we set out our reasons for proposing that large services and services that are multi-risk should track evidence of new kinds of illegal content on the service, and unusual increases in particular kinds of illegal content.
- 12.157 In instances where systems may need to deal with sudden harm events or spikes in illegal content, redeploying resource may draw resource away from another part of the system. Services which have contingency plans in place to ensure that illegal content across the system is dealt with expeditiously are more likely to protect their users appropriately. Hence it would be beneficial if services considered the potential for spikes in problematic and potentially illegal content when determining how to resource their content moderation functions.

### Language skills

- 12.158 Given the large number of languages that are spoken in the UK and the fact that some services may target specific communities of language speakers, content posted in many languages has the potential to cause harm to users in the UK. Where services consider what

---

<sup>75</sup> The Alan Turing Institute, 2022. [Tracking abuse on Twitter against football players in the 2021 – 22 Premier League Season](#). [accessed 21 August 2023].

<sup>76</sup> BSR, 2022. [BSR response to the 2022 Illegal Harms Call for Evidence](#).



language skills their content moderation teams may require to review potentially illegal content which could affect users in the UK, this is likely to reduce harm to deliver benefits.

- 12.159 The available evidence suggests that a range of stakeholders broadly agree with this hypothesis and that our proposal aligns with emerging industry practice. In response to the 2022 Illegal Harms Call for Evidence, a number of services and civil society organisations commented that moderation in different languages currently takes place or stressed the importance of doing so.
- 12.160 Through the 2022 Illegal Harms Call for Evidence, stakeholder engagement and other evidence, we are aware that several services already consider the language content is posted in and/or ensure they have the language expertise within their moderation systems to deal with it, using both humans and automated methods to do so.<sup>77</sup> In its response to the 2022 Illegal Harms Call for Evidence, [CONFIDENTIAL ✂].<sup>78</sup> [CONFIDENTIAL ✂].<sup>79</sup> Through stakeholder engagement, [CONFIDENTIAL ✂].<sup>80</sup> In a meeting, [CONFIDENTIAL ✂].<sup>81</sup> For example, in response to the 2022 Illegal Harms Call for Evidence, Glassdoor told us it uses proprietary technology to analyse all English and non-English language content.<sup>82</sup>
- 12.161 In response to the 2022 Illegal Harms Call for Evidence, a number of stakeholders, including BSR<sup>83</sup>, Chayn<sup>84</sup>, Glitch<sup>85</sup>, and Global Partners Digital<sup>86</sup>, stressed the importance of being able to moderate in different languages, as well as moderators having a knowledge of cultural context, to enable them to better understand the context relevant for the content being reviewed.<sup>87</sup>
- 12.162 There have been suggestions that many services 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.<sup>88</sup> A report by Demos, for example, noted that human

---

<sup>77</sup> 'The social media companies said they moderated content or provided fact-checks in many language s: more than 70 languages for TikTok, and more than 60 for Meta, which owns Facebook. YouTube said it had more than 20,000 people reviewing and removing misinformation, including in languages such as Mandarin and Spanish; TikTok said it had thousands. The companies declined to say how many employees were doing work in languages other than English.' The New York Times, 2022. . [accessed 3 August 2023].

<sup>78</sup> [CONFIDENTIAL ✂].

<sup>79</sup> [CONFIDENTIAL ✂].

<sup>80</sup> [CONFIDENTIAL ✂].

<sup>81</sup> [CONFIDENTIAL ✂].

<sup>82</sup> Glassdoor, 2022. [Glassdoor response to the 2022 Illegal Harms Call for Evidence](#).

<sup>83</sup> BSR, 2022. [BSR response to the 2022 Illegal Harms Call for Evidence](#).

<sup>84</sup> Chayn, 2022. [Chayn response to the 2022 Illegal Harms Call for Evidence](#).

<sup>85</sup> Glitch, 2022. [Glitch response to the 2022 Illegal Harms Call for Evidence](#).

<sup>86</sup> Global Partners Digital, 2022. [Global Partners Digital response to the 2022 Illegal Harms Call for Evidence](#).

<sup>87</sup> 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 3 August 2023].

<sup>88</sup> Avaaz, 2019. [Megaphone for Hate: Disinformation and hate speech on Facebook during Assam's citizenship count](#). [accessed 3 August 2023]; The Middle East Institute, 2020. [The flaws in the content moderation system: The Middle East case study](#). [accessed 3 August 2023]; New America, 2021. [Facebook's Content Moderation Language Barrier](#). [accessed 3 August 2023]; PBS, 2021. [Facebook's language gaps allow terrorist content and hate speech to thrive](#). [accessed 3 August 2023]; AACJ, 2022. [Fake News and the Growing Power of Asian American Voters: What this Means for 2022 Midterm Elections](#). [accessed 3 August 2023]; State of the Internet's Languages, 2022. [State of the Internet's Languages Report](#). [accessed 3 August 2023]; Global

moderators often have to make decisions about content in a language they do not understand.<sup>89</sup> Another report by the AI4Dignity Project – which focused on extreme speech specifically - noted that while companies are continuing to invest in natural language processing (NLP) models that cover a diversity of languages, existing AI models tend to cover large global languages, such as English, Spanish and Mandarin and may not cover smaller languages, noting this lack of linguistic diversity can result in harmful content being unidentified or misidentified.<sup>90</sup>

12.163 There is also the risk that a lack of language expertise in content moderation systems can lead to excessive moderation (or over-enforcement) of non-English or minority-language content, which poses a risk to freedom of expression. For example, Meta acknowledged that its current approach to moderating the Arabic word ‘shaheed’, which has multiple meaning but is often translated as ‘martyr’, may result in significant over-enforcement.<sup>91</sup>

12.164 The language expertise required to deal with the risk of harm in a particular language will likely differ from service to service based on a number of factors, including user base, content type and functionality. For this reason, we feel our Codes should not be prescriptive around what exact language expertise and resource is required on any service.

## Costs and risks

12.165 The cost of resourcing services’ content moderation systems adequately in line with this measure is likely to be substantial and ongoing. We expect it to vary by size of service and depend on the policies they develop and the nature and volume of illegal content on their service. In general, we would expect costs to be lower for smaller services and higher for larger services, everything else being equal. However, we are aware of a small service which needed to increase spending for online safety by several hundred thousand per annum to deal with problematic content on its service, some of which was illegal.<sup>92</sup> This illustrates the potentially substantial scale of the costs even small services may face where they are high risk.

12.166 The type of detection and review processes are likely to influence the magnitude of costs. Services have flexibility over the mix of human and automated content moderation they use:

- For example, automating content moderation processes (e.g. machine learning solutions for AI) require both one-off infrastructure investment, and different ICT professionals’ time. Larger services may be able to develop these in house, but the costs of doing so can be high.<sup>93</sup> Because of this, smaller services may outsource

---

Partners Digital, 2022. [Marginalised Languages and the Content Moderation Challenge](#). [accessed 3 August 2023]; Oversight Board, 2022. Oversight Board Annual Report 2021;

<sup>89</sup> Demos, 2020. [Everything in Moderation: Platforms, Communities and Users in a Healthy Online Environment](#). [accessed 3 August 2023].

<sup>90</sup> AI4Dignity, 2021. [Artificial Intelligence, Extreme Speech, and the Challenges of Online Content Moderation](#). [accessed 3 August 2023]; Columbia Journalism Review, 2021. [The challenges of global content moderation](#). [accessed 3 August 2023].

<sup>91</sup> Oversight Board, 2023. [Oversight Board announces a review of Meta's approach to the term "shaheed"](#). [accessed 3 August 2023].

<sup>92</sup> This is based on the increase in the number of content moderators that BitChute plans to put in place. This will increase to 21 content moderators, which we have used this to estimate the costs above.

<sup>93</sup> Ofcom, 2023. [Automated Content Classification \(ACC\) Systems](#) (a report for Ofcom by Winder.Ai), p.35. According to Winder.Ai, the cost of developing in-house AI software solutions by a small AI development team could exceed one million dollars. [accessed 26 September 2023].

development to a third party, or use off-the-shelf third-party solutions. Additionally, system updates, and licensing costs can be expensive and add to ongoing costs.

- If content moderation involves human moderators, resourcing costs will primarily depend on how many moderators are needed.<sup>94</sup> In addition, for content moderation to be effective human moderators may require specific tools to detect and review content and/or training (see from paragraph 12.175 below i.e. measure 6), but also a support ICT team. The service may also offer mental health support and other well-being benefits to its content moderators and other staff working on content moderation which would add to costs.
- Some services may require a separate review process for more complex illegal content cases, which may also require legal input.<sup>95</sup>

12.167 Services that already have policies and processes in place that are sufficient to meet this measure would not need to incur any additional costs, unless they wanted to withdraw those policies and processes.

## Rights impact

12.168 Our assessment of the rights impacts associated with having a content moderation function is set out above in relation to Measure 1 and our assessment of the implications of having performance targets is set out above in relation to Measure 3. We do not consider resourcing the function appropriately would have any additional impacts on those rights.

## Provisional conclusions

12.169 In view of the analysis above, we propose to recommend the following measure in our codes:

- a) Services which are large or multi-risk should resource their content moderation functions so as to give effect to their internal content policies and performance targets, having regard to: the propensity for external events to lead to a significant increase in demand for content moderation on the service; and the particular needs of its United Kingdom user base as identified in its risk assessment, in relation to languages.

12.170 Our analysis suggests that this measure could impose significant costs on services. However, for the reasons we explain above, we consider that if content moderation teams are not adequately resourced having regard to these factors this could significantly reduce their effectiveness. Given the importance of effective content moderation, this could give rise to very significant harm. While our proposed measure requires services to resource their content moderation functions to give effect to their performance targets, we do not propose to specify precisely what those performance targets are, which gives services some flexibility in precisely what they do. We therefore provisionally consider that this recommendation would be proportionate.

---

<sup>94</sup> The annual median content moderator earnings in the UK were £30,461 in 2022 (for further details see Annex 14), according to the ONS. Where content moderation occurs outside the UK, the US, Australasia or the European Union, the labour costs of human content moderators are likely to be lower than this.

<sup>95</sup> Google said in its responses to the 2022 Illegal Harms Call for Evidence that, 'Our legal removals team, comprising trained experts, reviews the report and determines whether to remove the content in accordance with applicable laws.' Google, 2022. [Google response to 2022 Illegal Harms Call for Evidence](#).

- 12.171 We are not at this point proposing extending the proposal to services that are not large and are not multi-risk. The amount and diversity of content such services need to moderate is likely to be materially lower and the benefits would therefore be materially smaller, making it questionable whether the potentially substantial costs of the measure were always justified for such services. Moreover, this measure is predicated on services having the internal content policies of our proposed Measure 2 above and the performance targets we propose in Measure 3, so it makes sense for this measure to apply to the same set of services as those proposed measures are recommended for.
- 12.172 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

## Measure 6: Services which are large or multi-risk should train people involved in content moderation and provide materials

---

- 12.173 As set out in relation to Measure 1, in order to comply with the Act, a service considering suspected illegal content should either make an illegal content judgment in relation to it, or, if it 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. It follows that the moderators carrying out this work need to know how to do whichever of those two things the service has chosen to do.
- 12.174 For small, low risk services which moderate little content, it may be possible to do this without training or additional written materials. But for services which are subject to Measure 2, we consider it very unlikely that it would be possible for moderators to give effect to content moderation policies without training and additional materials (such as: definitions and explanations around specific parts of the content moderation policy, enforcement guidelines, examples, and visuals of the tool or interface moderation staff will use to carry out their job). The extent of illegal content that larger and riskier services may face, as set out in paragraph 12.17 above, is far greater.

## Option(s) and Effectiveness

- 12.175 In this section, we are considering an option of recommending that services which have content moderation policies should ensure that people working in its content moderation process receive training and materials that enable them to moderate content in accordance with the other measures we propose in this chapter.
- 12.176 We know that many services already train their moderators and other relevant members of staff, or outsource to moderators and others who are trained<sup>96</sup>, to identify and remove illegal or violative content, as well as providing supporting materials to help them do so.<sup>97</sup>

---

<sup>96</sup> We know that outsourcing takes place in this sector: Morgan Lewis, 2023. [Emerging Market Trend: An Overview of Content Moderation Outsourcing](#). [accessed 25 September 2023]; NYU Stern Center for Business and Human Rights, 2020. [Who Moderates the Social Media Giants? A Call to End Outsourcing](#). [accessed 25 September 2023].

<sup>97</sup> 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 4 August 2023]. [CONFIDENTIAL ✕].

- 12.177 Several services told us they train their moderators to remove illegal (or violative) content and outlined (at a high-level) what kinds of training and support they receive.<sup>98</sup> For example, some services [CONFIDENTIAL X]<sup>99</sup> 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. These services also noted that they have on-going training, learning and development in place and that performance is assessed via exams.
- 12.178 Some services publicly outline what kinds of training and supporting materials they provide to their staff involved in content moderation. For example, Meta says its review teams “undergo extensive training to ensure that they have a strong grasp on our policies, the rationale behind our policies and how to apply our policies accurately.”<sup>100</sup> A number of services that use some form of community-reliant moderation have also developed moderation training and/or resources, including Discord<sup>101</sup>, Freecycle<sup>102</sup>, Nextdoor<sup>103</sup>, Reddit<sup>104</sup>, Twitch<sup>105</sup>, and WhatsApp.<sup>106</sup> However, it should be noted that the training and/or resources differ substantially from service to service and there appears to be no requirement that moderators complete this training before they begin moderating content.
- 12.179 In response to the 2022 Illegal Harms Call for Evidence, a number of civil society organisations, including 5Rights Foundation<sup>107</sup>, Carnegie UK<sup>108</sup>, the Center for Countering Digital Hate (CCDH)<sup>109</sup>, Refuge<sup>110</sup>, Glitch<sup>111</sup>, Global Partners Digital<sup>112</sup>, the NSPCC<sup>113</sup> and the Samaritans<sup>114</sup>, stressed the importance of training. For example, the NSPCC noted that for human moderators to be effective, “they should receive training so they can discharge their duties effectively and consistently”, while Global Partners Digital said that services should provide “extensive and regular training to moderators, on the detail and application of the respective terms of service”. The importance of training is also supported by broader academic and civil society literature and research.<sup>115</sup>

---

<sup>98</sup> [CONFIDENTIAL X].

<sup>99</sup> [CONFIDENTIAL X].

<sup>100</sup> Meta, 2022. [How review teams are trained](#). [accessed 4 August 2023].

<sup>101</sup> Discord, no date. [Discord Moderator Academy](#). [accessed 4 August 2023].

<sup>102</sup> Freecycle, no date. [Moderator Resources](#). [accessed 4 August 2023]; Freecycle, no date. [New Moderator Orientation](#). [accessed 4 August 2023].

<sup>103</sup> Nextdoor, no date. [About Review Team members and moderation](#). [accessed 4 August 2023].

<sup>104</sup> Reddit, no date. [Reddit mods](#). [accessed 4 August 2023].

<sup>105</sup> Twitch, no date. [Guide for Moderators](#). [accessed 4 August 2023].

<sup>106</sup> WhatsApp, no date. [101: Building a Safe Community](#). [accessed 4 August 2023].

<sup>107</sup> 5Rights Foundation, 2022. [5Rights Foundation response to 2022 Illegal Harms Call for Evidence](#). [accessed 4 August 2023].

<sup>108</sup> Carnegie UK, 2022. [Carnegie UK response to 2022 Illegal Harms Call for Evidence](#). [accessed 4 August 2023].

<sup>109</sup> Center for Countering Digital Hate (CCDH), 2022. [Center for Countering Digital Hate \(CCDH\) response to the 2022 Illegal Harms Call for Evidence](#). [accessed 4 August 2023].

<sup>110</sup> Refuge, 2022. [Refuge response to the 2022 Illegal Harms Call for Evidence](#). [accessed 4 August 2023].

<sup>111</sup> Glitch, 2022. [Glitch response to the 2022 Illegal Harms Call for Evidence](#). [accessed 4 August 2023].

<sup>112</sup> Global Partners Digital, 2022. [Global Partners Digital response to the 2022 Illegal Harms Call for Evidence](#). [accessed 4 August 2023].

<sup>113</sup> NSPCC, 2022. [NSPCC response to the 2022 Illegal Harms Call for Evidence](#). [accessed 4 August 2023].

<sup>114</sup> Samaritans, 2022. [Samaritans response to the 2022 Illegal Harms Call for Evidence](#). [accessed 4 August 2023].

<sup>115</sup> Cambridge Consultants, 2019. [Use of AI in Content Moderation](#). [accessed 3 August 2023]; Alan Turing Institute, 2021. [Understanding online hate: VSP Regulation and the broader context](#). [accessed 3 August 2023];

- 12.180 While services did not tell us exactly how often they trained staff involved in moderation, several did say they trained their staff regularly [CONFIDENTIAL ✕].<sup>116</sup>
- 12.181 Global Partners Digital told us that services should provide regular training to moderators, “on the detail and application of the respective terms of service and ensuring that moderators are aware of any changes made ahead of their implementation.”
- 12.182 The Trust & Safety Professional Association states on its website that before launching a policy change, staff involved in content moderation need to be trained on the change. Services may choose to carry out the training in a number of ways, either by giving it directly themselves, through external trainers, and/or via e-learning. Lastly, the Trust & Safety Professional Association said that minor policy or processes changes may take place via communication, for self-learning, rather than through training refreshers.<sup>117</sup>
- 12.183 Some stakeholders responding to the 2022 Illegal Harms Call for Evidence (Glitch and Global Partners Digital) also spoke about the importance of providing moderators with materials that support them in identifying and removing illegal content.
- 12.184 Specific materials provided to content moderators may include the content standards that fall under Measure 2 but also include 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 (i.e. the tool or interface moderation staff will use to carry out their job).<sup>118</sup> What is provided may vary depending on a number of factors, including, for example, the type of service, the type of content being moderated, and the local laws and regulations of the region where the service operates.
- 12.185 Based on the information above, we consider that training staff involved in moderation, as well as providing them with relevant materials, is beneficial, especially when compared to not training staff. Staff that have been trained on how to identify and remove illegal or violative content are more likely to be equipped with the knowledge and skills to do it when compared to those who are untrained.
- 12.186 We also think that staff involved in moderation who are trained regularly will have up-to-date knowledge of content moderation policies, as well as on the systems they are using to carry out their job.
- 12.187 There is no set best practice on how often training or supporting materials should be refreshed, and it may depend on a number of factors, including a person's role and performance. However, if moderators are trained on any major changes to policies or processes relating to content moderation, and provided with new or updated supporting materials, they are more likely to be able to give effect to them accurately and consistently.
- 12.188 We expect that the people 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, which includes, but is not limited to:

---

Brennan Center for Justice, 2021. [Double Standards in Social Media Content Moderation](#). [accessed 3 August 2023].

<sup>116</sup> [CONFIDENTIAL ✕].

<sup>117</sup> Trust & Safety Professional Association, no date. [Setting Up A Content Moderator for Success](#). [accessed 4 August 2023].

<sup>118</sup> The Guardian, 2017. [Revealed: Facebook's internal rulebook on sex, terrorism and violence](#). [accessed 4 August 2023]; Trust & Safety Professional Association, no date. [Setting Up A Content Moderator for Success](#). [accessed 4 August 2023].



Trust and Safety staff; quality assurance and compliance staff; subject matter experts; lawyers and other legal staff; risk management staff; operations staff; engineers; and developers.

- 12.189 We are aware that many services use volunteers to help them moderate content (sometimes referred to as ‘community-reliant’ moderation), which can have both benefits and drawbacks for services and the safety of users. We also know that many services that use voluntary moderators have developed training and/or resources to support community moderation. However, there would be a significant extra cost burden for services if we were to extend the measure to volunteer moderators, and at this early stage we are not in a position to predict the possible impact on services’ businesses of making such a recommendation. This option does not include that volunteers should be trained notwithstanding that we recognise that this could give rise to risks for users. We note that our evidence suggests that the majority of services will have some paid staff that deal with moderation alongside using voluntary moderators.<sup>119</sup>
- 12.190 We therefore consider that users would be better protected from harm if we recommend that a U2U service which has a moderation policy should ensure that people working in its moderation process (other than volunteers) receive training and materials that enable them to moderate in accordance with the other measures above.
- 12.191 We do not consider that it would be appropriate to specify in Codes how often materials should be revised or training should be redelivered. A service which failed to refresh training and materials following any major changes to policies or processes relating to content moderation that is to do with suspected illegal content would not be enabling its moderators to moderate content in accordance with Measures 1 to 5 above.

### Possible factors to consider in the training of staff involved in content moderation and supporting materials

- 12.192 As set out above, we consider that generally speaking services are best placed at present to determine what is appropriate for their services in terms of the detail of their training and materials. However, services which do not have regard to certain factors are unlikely to protect users properly. We therefore consider below whether to specify in Codes that in preparing and delivering content moderation training and materials, services should have regard to particular matters.
- 12.193 Risk assessment and information pertaining to the tracking of signals of emerging harm - A service's risk assessment will be one of the key sources of information telling a service what risk of illegal content they have on their platform and will form the basis for internal content policies (see Measure 2). As moderators should be focused on enforcing the internal content policies, training should also be informed by the most recent illegal content risk assessment. In Chapter 8, we are also consulting on a proposed recommendation that services should track signals of emerging harm. If, following consultation, we remain of the view we should recommend this, this information would be one of the key sources of information about how illegal content manifests and it is therefore crucial services use this to inform their content moderation training and supporting materials.

---

<sup>119</sup> New America, 2019. [Everything in Moderation – Case Study: Reddit](#). [accessed 30 August 2023]; Nextdoor, no date. [About moderation](#). [accessed 30 August 2023]; Twitch, no date. [Moderation on Twitch](#). [accessed 30 August 2023]; [CONFIDENTIAL ✕].



- 12.194 Remedying gaps in moderation staff’s understanding of specific harms – In response to the 2022 Illegal Harms Call for Evidence, 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.<sup>120</sup> 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.<sup>121</sup> We also know that many services, particularly larger ones, give their staff involved in moderation specialist training and materials in particular areas, including illegal harms, other harms, freedom of expression, and user rights.<sup>122</sup>
- 12.195 Several civil society organisations recommended specialist training on specific harm areas, including, tech abuse and gender-based violence (Glitch<sup>123</sup> and Refuge<sup>124</sup>); child safeguarding, risks to children, and knowledge of child development (5Rights Foundation<sup>125</sup> and NSPCC<sup>126</sup>); and awareness of learning disabilities (MENCAP<sup>127</sup>). Global Partners Digital also stressed the importance of training moderators in the potential impact to users’ rights and freedom of expression.<sup>128</sup>
- 12.196 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. For example, although some CSAM can be easily identified as illegal content, there are many exceptions to this. For example, it can be difficult for content moderators to determine whether an image depicts a person who is under or over 18. If training and materials are given to moderators where a service has identified a gap in moderators’ understanding of a specific harm, and where they deem there to be a specific risk, this should improve outcomes for users.

## Other issues to note

- 12.197 A number of civil society respondents to the 2022 Illegal Harms Call for Evidence stressed the importance of supporting the wellbeing of staff involved in content moderation,

---

<sup>120</sup> OnlyFans, 2022. [OnlyFans response to the 2022 Illegal Harms Call for Evidence](#).

<sup>121</sup> Nextdoor, 2022. [Nextdoor response to the 2022 Illegal Harms Call for Evidence](#).

<sup>122</sup> [CONFIDENTIAL ✕].

<sup>123</sup> Glitch said there should ‘comprehensive’ training for moderators on “online gender-based violence and different tactics of online abuse, and how abuse specifically targets women, Black and minoritised communities and users with intersecting identities”. Glitch, 2022. [Glitch response to the 2022 Illegal Harms Call for Evidence](#).

<sup>124</sup> Refuge noted that moderators should be fully trained in identifying and responding to different types of tech abuse and other forms of VAWG, because they say to the untrained eye, tech abuse can often be hard to recognise. Refuge, 2022. [Refuge response to the 2022 Illegal Harms Call for Evidence](#).

<sup>125</sup> 5Rights Foundation commented that human moderators should receive training in how to identify risks to child safety, “including knowledge of risks to different groups of children and the full range of content and activity that is illegal or might be harmful to a child. This also includes knowledge of the stages of child development and awareness of how children’s capacities, vulnerabilities and behaviour change as they grow.” 5Rights Foundation, 2022. [5Rights Foundation response to the 2022 Illegal Harms Call for Evidence](#).

<sup>126</sup> The NSPCC said that moderators looking at CSA content and activities should be trained in moderation and safeguarding. NSPCC, 2022. [NSPCC response to the 2022 Illegal Harms Call for Evidence](#).

<sup>127</sup> MENCAP said that to moderate content more accurately, there should be “awareness training to moderators on learning disability as well as other groups deemed more likely to be subjected to online harms and illegal content.” MENCAP, 2022. [MENCAP response to the 2022 Illegal Harms Call for Evidence](#).

<sup>128</sup> Global Partners Digital, 2022. [Global Partners Digital response to the 2022 Illegal Harms Call for Evidence](#).

including Carnegie UK, Chayn, Glitch, and Global Partners Digital. This was also noted by some services [CONFIDENTIAL X].<sup>129</sup> For example, Glitch said that content moderators “should work in holistic environments which appropriately support their wellbeing, proportionate to the level of upsetting and harmful material they are moderating”. Global Partners Digital noted that adequate financial, emotional and psychological support is “vital to reduce turnover and burnout in content moderation teams, which limits institutional knowledge and consistency between decisions and lowers the overall accuracy of the content moderation systems”.

- 12.198 Research suggests that human content moderation has the potential to cause significant impacts on the wellbeing of staff members, including secondary trauma, altered psychological wellness, and burnout.<sup>130</sup> Some platforms offer controls to moderators when reviewing content, such as applying blurring or audio removal, though this is not universal.<sup>131</sup> Some platforms also have wellbeing support in place for moderators such as counselling and mental health support, such as [CONFIDENTIAL X].<sup>132</sup>
- 12.199 We recognise the significant impact that human moderation of content can have on the wellbeing of an individual and the importance of providing appropriate supervision and support in this area. However, the wellbeing of content moderators would only be relevant to our remit if it impacted on user safety. We welcome evidence from stakeholders on this, to which we would have regard in planning our work on future iterations of our Codes.

## Costs and risks

- 12.200 The main factors driving the cost of the training would be the number of staff to be trained and the duration of the training. The duration of the training needed will tend to be longer the more complex and diverse the range of possible illegal content on a service. For the duration of the training, we assume a range of two to six weeks for someone having this training for the first time.<sup>133</sup> Based on this duration and a range for pay, we estimate that the costs of providing training for one new content moderator could be between £2,500 and £15,000, and for a new software engineer between £3,500 and £21,000.<sup>134</sup> If content moderators are based in countries with lower labour costs than the UK, then the lower end

---

<sup>129</sup> [CONFIDENTIAL X].

<sup>130</sup> Steiger, M., Bharucha, J.T., Venkatagiri, S., Martin J. Riedl, J.M., and Lease, M., 2021. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.

<sup>131</sup> Spence, R., DeMarco, J., and Martellozzo, E., 2022. [Invisible workers, hidden dangers](#). [accessed: 14 September 2023].

<sup>132</sup> [CONFIDENTIAL X].

<sup>133</sup> 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 services may also provide some refresher training to its employees from time to time, which is likely to vary by service and depend on several factors including the individual and their role.

<sup>134</sup> This is based on our assumptions on wage rates set out in Annex 14. 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 staff being trained accounted for about half of all training expenditure in 2019, although this varies by size of the firm and sector. We assume this excludes the 22% uplift that we have elsewhere assumed for non-wage labour costs, so we have not also increased these wages by 22%. Source: Department for Education (DfE), 2020. [Employer Skills Survey 2019: Training and Workforce Development – research report](#), pp38 and 40. [accessed: 14 September 2023].

of the wage range we have assumed will overstate the costs. These costs may also vary depending on whether the training is by in-house staff or by an external provider.<sup>135</sup>

- 12.201 In addition to these costs of training new content moderators and software engineers, there will also be some ongoing costs for refresher training and training in new harms on the services. We expect the annual costs of these to be lower.
- 12.202 All else being equal, smaller services will have less content to review, smaller content moderator teams and therefore lower costs.<sup>136</sup> While costs for smaller (and larger) services will scale with the risk of harm, this will come with a commensurate benefit. In broad terms, we would expect costs to vary with the potential benefits, in the sense that more content moderators will be needed, the more illegal content tends to be on a service.
- 12.203 As discussed in paragraph 12.189 – voluntary content moderators, many services use volunteers to help them moderate content. For some services, this can involve large numbers of such volunteers.<sup>137</sup> Given this and the costs of training per content moderator, we are not recommending a requirement on services to train and provide materials to voluntary moderators.
- 12.204 There may also be costs involved with any additional materials for content moderators which were not used in the training. We do not anticipate the costs of preparing and producing such materials to add much to the costs of the training.

## Rights impacts

### Freedom of expression

- 12.205 We would not expect this option to have any negative impacts on the rights to freedom of expression. As several respondents to the 2022 Illegal Harms Call for Evidence noted, training enables those involved in content moderation to make better decisions [CONFIDENTIAL ✕].<sup>138</sup> Training also enables staff involved in moderation to have a better understanding of borderline content, i.e. content where it can be difficult to determine whether it is legal or illegal.

---

<sup>135</sup> Based on FCA's research, all large firms and 40% of medium firms are assumed to have in-house training departments. Source: Financial Conduct Authority ("FCA"), 2018. [How we analyse the costs and benefits of our policies](#), p.44. [accessed: 14 September 2023].

<sup>136</sup> For example, the cost of Mumsnet training its Community team of 14 freelance moderators and two staff members, would be considerably different from Meta that employs 15,000 content reviewers around the world, although the benefits would also be different (Mumsnet with ~ eight million unique users monthly and Meta with 3.74bn monthly users across its platforms globally in December 2022). Mumsnet, 2022, and Meta, 2022, responses to the 2022 Illegal Harms Call for Evidence. Meta, 2023, [Meta Reports Fourth Quarter and Full Year 2022 Results](#). [accessed 17 September 2023].

<sup>137</sup> A 2022 study from academics at Northwestern University and the University of Minnesota Twin Cities said there were 21,522 active Reddit community moderators. Li, H., Hecht, B. and Chancellor, S., 2022. Measuring the Monetary Value of Online Volunteer Work. In: Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 596-606. In its annual Transparency Report for 2022, Nextdoor said it had 210,900 volunteer community moderators. Nextdoor, 2023. [Nextdoor publishes second annual Transparency Report, revealing record low levels of harmful content reported on the platform](#). [accessed 30 August 2023].

<sup>138</sup> Wikimedia, 2022. [Wikimedia response to the 2022 Illegal Harms Call for Evidence](#). [CONFIDENTIAL ✕].

## Privacy

- 12.206 Services would need to comply with privacy and data protection laws in relation to any items of content they use in their training and other materials.
- 12.207 We consider that the training of moderators would be a further safeguard for users' privacy, against the possibility that services may incorrectly report detected illegal content to reporting authorities.

## Provisional conclusion

- 12.208 As set out above, this option is linked to and would be effective for those services which have search moderation policies in compliance with Measure 2. It follows that it should only be considered for those services – i.e. large services and multi-risk services.<sup>139</sup>
- 12.209 We recognise that the additional costs may be significant for some services. However, we also consider the benefits of this measure are likely to be high. This is because content moderator training is important in effectively implementing a service's content moderation policies to reduce harm and comply with its online safety duties. Well-trained and prepared content moderators are more likely to be able to identify content in accordance with Measure 1 and the service's content standards (under Measure 2), and to apply the correct treatment to it, reducing the harms that result from that. As the number of content moderators that need training is likely to depend on the volume of content that needs to be assessed, the costs of this measure are likely to scale with the benefits. As such, this measure is likely to be proportionate for services which identify significant risks to users.
- 12.210 We consider this to be the case for both large and smaller services. Training costs are likely to depend primarily on the number of people that need to be trained. Everything else being equal, smaller services are likely to have smaller volumes of content, and fewer content moderators as a result. This means the costs for smaller services will be correspondingly lower than for large services.
- 12.211 For these reasons, our provisional view is that it is proportionate to recommend this measure to large services and to multi risk services.
- 12.212 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.
- 12.213 The full text of our proposed measure, covering each of the factors outlined in our discussion above, can be found in our proposed Code of practice, Annex 7, Recommendation 4F.

---

<sup>139</sup> See paragraphs 11.43 to 11.46 for how we propose to define multi-risk.

# 13. Search moderation

## What is this chapter about?

This chapter discusses the steps we expect search services to take to moderate search content which they index.

## What are we proposing?

We are making the following proposal for all search services:

- **Have systems or processes designed to deindex or downrank illegal content of which it is aware, that may appear in search results.** In considering whether to deindex or downrank the content concerned, services should have regard to the following factors: (i) the prevalence of illegal content hosted by the interested person; (ii) the interests of users in receiving any lawful material that would be affected; and (iii) the severity of harmfulness of the content, including whether or not the content is priority illegal content.

We are making the following proposals for all large general search services and any other multi-risk search services:

- **Set and record internal content policies. These should set out rules, standards and guidelines about: what content is allowed and not allowed on the service, and how policies should be operationalised and enforced. In doing so, services should have regard to its risk assessment and signals of emerging illegal harm.**
- **Set and record performance targets for its search moderation functions and measure and monitor its performance against these targets.** These should include the time that illegal content remains on service before it is deindexed or downranked, and the accuracy of decision making. When setting targets, services should balance the desirability of deindexing or downranking illegal content swiftly against the need to make accurate moderation decisions.
- **Prepare and apply a policy about the prioritisation of content for review. This policy should have regard to at least the following factors: virality of content, potential severity of content, and the likelihood that content is illegal, including whether it has been flagged by a trusted flagger.**
- **Resource its search moderation function so as to give effect to their internal content policies and performance targets.** In doing so, it should have regard to the propensity for increases in demand for search moderation caused by external events. When deciding how to resource their functions services should consider the particular needs of its UK user base, in relation to languages.
- **Ensure people working in search moderation receive training and materials that enable them to moderate content effectively.**

## Why are we proposing this?

In order to protect their users, search services are required to take proportionate steps to minimise the risk of individuals encountering illegal content in searches, for example by deindexing or

downranking it. We refer to these activities as search moderation. Effective search moderation plays an important role in protecting users from harm associated with illegal content.

Whilst search services will always need to take action where they have reasonable grounds to infer that search content such as a webpage contains illegal content, it may not always be appropriate to deindex it. For example, if that webpage contained only a small amount of less severe illegal content and a large volume of valuable lawful content, it may be more appropriate to downrank the webpage instead. Conversely, where a webpage contains the most severe forms of illegal content, deindexing is likely to be more appropriate. We therefore propose to give search services a degree of flexibility as to whether to deindex or downrank webpages containing illegal content, depending on the specific context.

Our analysis suggests that harm to users will be reduced where search services set content policies, resource and train their search moderation teams adequately and take into account the likely severity of content and the frequency with which it is searched when deciding what potentially harmful search content to prioritise for review. Given the diverse range of services in scope of the new regulations, a one-size-fits-all approach to search moderation would not be appropriate. Instead of making very specific and prescriptive proposals about search moderation, we are therefore consulting on a relatively high-level set of recommendations which would allow services considerable flexibility about how to set up their search moderation functions.

We have focussed the most onerous proposals in this area on large general search services and any other search services which are 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 search services to carry them out in a way which is cost-effective and proportionate for them.

We recognise that search services often use a combination of automated tools and human review to moderate search content. The proposals in this chapter are not prescriptive about the balance services should strike between human and automated review of content and would not require services to use automated tools to review content. Where we have made specific recommendations about automated review of search content we consider these separately and in more detail in a later chapter.

## What input do we want from stakeholders?

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

## Introduction

---

- 13.1 In Chapter 12 we considered proposals in relation to content moderation on U2U services. In this chapter we consider what steps search services should take by way of moderation.
- 13.2 Under the Act, a ‘search service’ is defined as “an internet service that is, or includes a search engine” and a search engine “includes a service or functionality which enables a person to search some websites or databases” but “does not include a service which enables a person to search just one website or database”.<sup>140</sup>

---

<sup>140</sup> See section 3(5) and 229(1) of the Act

- 13.3 As set out in Chapter 11, we distinguish between the following types of search services: general search services (which enable users to search the web by inputting search queries on any topic) and vertical search services (which focus only on a specific topic or genre of content). Within general search we also distinguish between services that only rely on their own indexing and those which contract to obtain search results (which we call downstream general search services). A longer description of each of these types of service can be found in paragraph 11.65.
- 13.4 Broadly, the Act requires that a search provider must take measures relating to the design and operation of its service to effectively mitigate and manage the risks of harm identified in the risk assessment. It must also operate the service in a way that minimises the risk of individuals encountering search content that is illegal content. These duties apply across the whole service but, where proportionate, the provider must adopt measures in particular areas, including in relation to functionalities and content prioritisation.<sup>141</sup>
- 13.5 In practice, this means that a service provider is expected to minimise the risk of individuals encountering illegal content in or via its search results by moderating search content on its service. It is important to recognise that content is to be treated as ‘encountered via’ search results where it is encountered as a consequence of interacting with results (for example by clicking on them).<sup>142</sup> This means that search content includes content on a webpage that can be accessed by interacting with search results. The safety duties, and the measures we recommend for the purposes of complying with them below, should be considered in this context.
- 13.6 As with content moderation in U2U services, the exact methods and techniques for doing this may vary between services and there is no ‘one-size-fits-all’ approach to moderating search results. We recognise that moderation systems and processes differ from service to service and are designed to meet specific needs and contexts, and the measures we recommend below are intended to reflect this.
- 13.7 As discussed below, there are different ways that a search service might choose to moderate search content for the purposes of complying with its duties. This may include deindexing, downranking or other forms of prioritisation. For the purposes of this chapter, references to search ‘moderation’ (and associated expressions) should be construed as including all such actions. The recommendations set out in this chapter engage the rights of users and providers in a similar way to the recommendations set out in Chapter 12. One key distinction is that the person who is responsible for the content may have no relationship whatsoever with the search service. There is nevertheless a need to take account of the rights of those responsible for websites or databases that are capable of being searched by providers’ search engines (so far as they are based in the United Kingdom); this group is referred to in the Act, and below, as ‘interested persons’.<sup>143</sup>

---

<sup>141</sup> See section 27(4) of the Act

<sup>142</sup> This does not extend to subsequent interactions with anything other than a search result. Source: section 57(5) of the Act.

<sup>143</sup> For the definition of “interested persons” see section 227(7) of the Act.



## Harms the measures seek to address

- 13.8 There is evidence that general search services can be used to access content related to a wide range of offences, including, amongst other things, terrorism, hate, extreme pornography, CSAM (Child Sexual Abuse Material) and fraud.<sup>144</sup>
- 13.9 Under section 27 of the Act, regulated search services must take steps to reduce the risk of harm to users identified in their most recent illegal content risk assessment, and to minimise the risk of individuals encountering both search content that is priority illegal content and relevant non-priority illegal content of which they are aware (section 27(2) and (3)).
- 13.10 These duties differ from the duties applicable to U2U services. There is no duty to take down illegal content swiftly or minimise how long it is present (because search services do not control the content). Nor is there a duty to take proportionate steps to prevent users from encountering search content that is illegal content.
- 13.11 The safety duties differ as between search services and U2U services in another respect: U2U services must, where it is proportionate, take or use measures in the area of ‘content moderation’ where it is proportionate. While there is no such express duty for search services, they are required to take or use ‘content prioritisation’ measures in seeking to comply with the safety duties. As such, in order for search services to fulfil their duties under the Act, it is clear that they will at least need to be able to consider whether or not search content is illegal content.
- 13.12 In addition, like U2U services, search services are required to enable their users to make complaints about illegal content, albeit that for search services these are complaints about search content that is illegal content (sections 31 and 32). They must take ‘appropriate action’ in response to such complaints (section 32).
- 13.13 It is difficult to see what action could be ‘appropriate’ in response to such complaints, absent a capacity to consider whether or not action should be taken against the content in question. We note, in particular, that section 32(4)(c) of the Act assumes that search services may take or use measures in order to comply with a duty set out in section 27, in a way that results in content no longer appearing in search results or being given a lower priority in search results. We consider the appropriate action in response to search complaints further in Chapter 16, from paragraph 16.157.
- 13.14 Overall, therefore, while the Act does not expressly require search services to have a proportionate ‘content moderation’ function, the effect of sections 27 and 32 is that they need a function capable of making judgments about whether search content should be treated as illegal content. We are calling this function ‘search moderation’.
- 13.15 As for the U2U services we considered in Chapter 12, the harms we consider in this chapter potentially arise on all search services, but to different degrees. We have no evidence of harms arising on vertical search services. Some smaller general search services may not have very much search moderation to do (e.g. because they receive hardly any complaints). By contrast, larger general search services may face significant challenges in terms of the volumes and diverse nature of the search content they need to moderate, giving risk to questions about how to prioritise content for review, achieve

---

<sup>144</sup> Register of Risks, Volume 2, Chapter 6 Part 2: Search services.

consistency, quality and timeliness of decision-making, and plan their deployment of search moderation resourcing so as to secure that users are appropriately protected.

## Proposed approach

---

### How we have approached the provisions in Codes

- 13.16 In light of the analysis above, we consider that it is important to include recommendations about search moderation in our Codes. As with content moderation for U2U services, we have considered three potential approaches to drafting these measures for search services:
- i) **Approach 1** - specify in detail how services should configure their search moderation systems and processes;
  - ii) **Approach 2** - specify in detail the outcomes search moderation systems and processes should achieve (i.e. setting detailed KPIs), but leave the design to services; or,
  - iii) **Approach 3** - require services to operate a search moderation system and (where relevant) set out the factors to which they should have regard when designing their content moderation systems and processes.
- 13.17 Evidence is limited on how search services resource their moderation systems and processes. We know that some larger search services use a combination of human and automated moderation to minimise the risk of users encountering illegal content.<sup>145</sup> In its response to the 2022 Illegal Harms Call for Evidence, Google told us it is constantly hiring new people dedicated to safety policy, as well as investing in new technology to help it tackle illegal and harmful content at scale. It also told us it has “*long invested in the most effective automated systems for protecting users from harmful content and [has] developed effective automated detection tools*”.<sup>146</sup> Google, for example, has Trust and Safety staff working across its search product to help tackle both harmful and illegal content, as well as employing content moderators (both internally and externally, i.e. contractors) to review and remove harmful and legal content.<sup>147</sup> Google Search also uses automated systems to “*help protect against objectionable material*”.<sup>148</sup> Similarly, Microsoft Bing outlines on its website that it uses ‘complex algorithms’ to generate search results and may automatically moderate content if it is potentially harmful, as well as using humans to review reports of potential content violations.<sup>149</sup> Yahoo notes that it uses ‘content moderation tools’.<sup>150</sup>
- 13.18 However, we also know that some small search services rely solely on human resource to deprioritise or deindex illegal content in search results. In its response to the 2022 Illegal Harms Call for Evidence, [CONFIDENTIAL X].<sup>151</sup>

---

<sup>145</sup> Trust and Safety Professional Association, no date. [The Purpose and Role of T&S Teams](#). [accessed 16 August 2023]; Cambridge Consultants, 2019. [Use of AI in Content Moderation](#). [accessed 3 August 2023]; Google, 2019. [Meet the teams keeping our corner of the internet safer](#). [accessed 16 August 2023].

<sup>146</sup> [Google response to the 2022 Illegal Harms Call for Evidence](#). [accessed 27 September 2023].

<sup>147</sup> Google, 2019. [Meet the teams keeping our corner of the internet safer](#). [accessed 16 August 2023].

<sup>148</sup> Google, no date. [Content policies for Google Search](#). [accessed 17 August 2023].

<sup>149</sup> Microsoft Bing, no date. [Bing Webmaster Guidelines](#). [accessed 16 August 2023]; Microsoft, 2023. [How Bing delivers search results](#). [accessed 16 August 2023].

<sup>150</sup> Yahoo, 2021. [Yahoo Community Guidelines](#). [accessed 16 August 2023].

<sup>151</sup> [CONFIDENTIAL X].

- 13.19 Automated content moderation (ACM) tools also are a resource that can be deployed across systems to tackle illegal and/or harmful content. Due to the complexities of harms, and the intrinsic limitations of individual automated content moderation technologies, it is often the case that services will use automated tools in conjunction, layering one measure on top of another, as well as other signals, to assess with sufficient confidence whether a piece of content is violative or illegal and should be removed. In its response to the 2022 Illegal Harms Call for Evidence, Google said it has built a range of products, tools and approaches across its different services that ensure users can have a safe experience.<sup>152</sup>
- 13.20 Services may also use specific resources to tackle certain harms. For example, Google says it uses hash matching and artificial intelligence technologies to proactively identify CSAM, constantly updates its algorithms to tackle this evolving threat, and uses teams of *“highly specialized and trained content reviewers and subject matter experts”*.<sup>153</sup> Similarly, for Bing, Microsoft says it *“works to prevent CSEAI [child sexual exploitation and abuse imagery] from entering the Bing search index by leveraging block lists of sites containing CSEAI identified by credible agencies, and through PhotoDNA scanning of the index and visual search references when users upload images on one of Bing’s hosted features such as visual search”*.<sup>154</sup>
- 13.21 Several civil society organisations, including Business for Social Responsibility (BSR), Carnegie UK and Global Partners Digital, stressed the importance of ensuring there is sufficient coverage of human content moderators, both in hours covered by shifts and the number of employees, to allow moderators adequate time to review each piece of content.
- 13.22 We provisionally consider that there is no ‘one size fits all’ approach to resourcing a search moderation system or to defining the outcomes it should achieve. There is significant diversity and innovation in moderation processes.
- 13.23 Our analysis of the benefits and drawbacks of each of the approaches set out above is reflective of the analysis already carried out in the U2U services content moderation chapter (see paragraphs 12.18 to 12.44). For these reasons, we will not repeat that analysis here. With the exception of the measure set out in Chapter 15, it is our provisional view that the **third approach would be most appropriate for search services**. This is because:
- a) we do not currently have enough evidence to specify in detail how search services should configure their search moderation systems and processes (approach 1) or specify in detail the outcomes search moderation systems and processes should achieve (approach 2). Taking a prescriptive and specific approach at this stage would therefore give rise to a substantial risk of regulatory failure and unforeseen consequences. It could lead to significant disruption in the sector and potentially to increased, rather than decreased, harm to users; and
  - b) it would allow search services greater flexibility on how to behave to achieve compliance and allow services to comply with the measures in ways that may be

---

<sup>152</sup> [Google response to the 2022 Illegal Harms Call for Evidence](#). [accessed 27 September 2023].

<sup>153</sup> Google, no date. [Fighting child sexual abuse online](#). [accessed 17 August 2023]; Google, 2022. [How we detect, remove and report child sexual abuse material](#). [accessed 17 August 2023].

<sup>154</sup> Microsoft, no date. [Digital Safety Content Report](#). [accessed 17 August 2023].

more proportionate and cost effective for them, while also still setting out the important factors that services should take into account. This is particularly beneficial in this context given the diverse range of services in scope of the Act and the fast-moving pace of technological development.

- 13.24 However, where we are able to identify with a good degree of confidence particular things that we think services should be doing, or particular outcomes that we think they should be achieving, then we propose to be more specific in our Codes. See Chapter 15 Automated Content Moderation - Search.

## Relationship between publicly available statements and illegal content judgments in search moderation

- 13.25 As set out in relation to U2U services in Chapter 12, we recognise that many service providers will have designed their publicly available statements to comply with existing laws in multiple jurisdictions and their own commercial needs, and these may be effective to secure that illegal content is dealt with.
- 13.26 The Act allows for service providers to have different terms of service for UK users when compared to users elsewhere in the world. In practice, where the Act requires content to be deindexed or downranked, this means deindexed or downranked for UK users.
- 13.27 Services may become aware of suspected 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 16. In the same chapter, we also consider whether to propose a means for entities with appropriate expertise and information ('trusted flaggers') to report suspected illegal content to services. In Chapter 15, we identify the automated content moderation (ACM) technology we propose to recommend with a view to identifying further illegal content. Services may choose to use other kinds of technology or human content moderators in order to identify suspected illegal content as defined in the Act.

## Search moderation systems

---

### Options

- 13.28 We have considered the case for recommending the following measures relating to search moderation in our codes:
- a) Measure 1(a): All search services should deindex URLs where there are reasonable grounds to infer they contain illegal content;
  - b) Measure 1(b): All search services should either deindex or, alternatively, down rank URLs where there are reasonable grounds to infer they contain illegal content. When deciding whether to deindex or downrank, they should have regard to at least the following factors: (i) the prevalence of illegal content hosted by the interested person on the URL <sup>155</sup>; (ii) the interests of users in receiving any lawful material that would be affected; and (iii) the severity of harmfulness of the content, including whether or not the content is priority illegal content.

---

<sup>155</sup> For URLs that may contain CSAM, we note that services will need to rely on expert organisations, given the legal risks of reviewing CSAM.

- c) Measure 2: Large general search services or multi-risk search services should set internal content policies having regard to at least the findings of their risk assessment and any evidence of emerging harms on their service.
- d) Measure 3: Large general search services or multi-risk search services should set performance targets for their search moderation functions and track whether they are meeting these. These should include targets for both how quickly URLs containing illegal content are deindexed or downranked and for the accuracy of search moderation decisions. When setting targets, services should balance the need to deindex or downrank URLs containing illegal content swiftly against the need to make accurate moderation decisions. They should measure their performance against their targets.
- e) Measure 4: Large general search services or multi-risk search services should have and apply policies on prioritising search content for review. In setting the policy, the provider should have regard to at least the following factors: (1) how frequently search requests for the search content are made ; the potential severity of the search content: including whether the content is suspected to be priority illegal content and the provider's risk assessment; (3) the likelihood that the search content is illegal content, including whether it has been reported by a trusted flagger.
- f) Measure 5: Large general search services or multi-risk search services should resource their search moderation functions so as to give effect to their internal content policies and performance targets, having regard to at least: (i) the propensity for external events to lead to a significant increase in demand for search moderation on the service; and (ii) the particular needs of its United Kingdom user base as identified in its risk assessment, in relation to languages.
- g) Measure 6: Large general search services or multi-risk search services should ensure their search moderation teams are appropriately trained.

13.29 We set out our assessment of the case for these measures below.

## Measure 1: Having in place moderation systems or processes designed to deindex or downrank illegal content

---

13.30 The safety duties in the Act in effect require that search services must have in place systems or processes to moderate search content that is illegal content. The Act requires that search service take steps to minimise the risk of individuals encountering search content that is illegal content. While search services may be unable to take down illegal content as U2U services might, they may nonetheless take other steps to minimise the extent to which users encounter it. The two main ways they have to do this are deindexing URLs containing potentially illegal content or downranking them. We explore these concepts below and consider two options:

- a) Recommending that search services always deindex URLs when there are reasonable grounds to infer the contain illegal content; or
- b) Recommending that all search services should either deindex or, alternatively, down rank URLs where there are reasonable grounds to infer they contain illegal content. When deciding whether to deindex or downrank, they should have regard to at least

the following factors: (i) the prevalence of illegal content hosted by the interested person on the URL; (ii) the interests of users in receiving any lawful material that would be affected; and (iii) the severity of harmfulness of the content, including whether or not the content is priority illegal content.

## Deindexing or downranking illegal content

- 13.31 **Deindexing** (also known as delisting) content is typically understood to involve the removal of links to webpages, preventing those links from being accessed through the search results. It can be implemented at the URL level (i.e. deindexing individual webpages) or at the domain level (i.e. deindexing entire websites). For example, a search service may choose to deindex the URL of a webpage that contains illegal content, or it may choose to deindex an entire website at domain level if it thinks there is illegal content present sitewide.
- 13.32 Whilst deindexed content may still be accessed via open web, social media platforms and searches on other platforms, deindexing remains the principal means by which general search services may control the visibility of, and access to, content presented to users in its search results. As such, it should be regarded as a key tool for providers for the purposes of seeking to comply with the safety duties.
- 13.33 We understand that deindexing is commonly used by general search services:
- a) In response to our 2022 Illegal Harms Call for Evidence, Google provided information on its policies for delisting (deindexing). Google’s content policies state that it delists *“certain personal information that creates a significant risk of identity theft, financial fraud, or other specific harm, non-consensual explicit imagery (NCEI), search results that lead to child sexual abuse imagery or material that appears to victimise, endanger, or otherwise exploit children”* and that it either delists or demotes spam, defined as *“results that exhibit deceptive or manipulative behaviour designed to deceive users or game our search systems”*.<sup>156</sup>
  - b) Microsoft Bing’s webmaster guidelines describe the approach to content ranking on Bing, and stress that de-indexing is restricted to *“a narrow set of circumstances and conditions to avoid restricting Bing users’ access to relevant information”*.<sup>157</sup>
- 13.34 A search service may also **downrank** content by altering the ranking algorithm to ensure that a particular piece of content appears lower in the search results and is therefore less discoverable to users.
- 13.35 In many cases, providers already deindex or downrank content that breaches their policies.
- 13.36 We carefully considered whether there was a case for recommending that providers should always deindex, rather than downrank, search content that is illegal content where they come across it on their service. The argument might be made that, given it is illegal, downranking is not enough, because users can still access downranked content via the service. However, we provisionally consider that, at least at this stage, such a recommendation cannot be justified on proportionality grounds.

---

<sup>156</sup> [Google response](#) to 2022 Ofcom Call for evidence: First phase of online safety regulation. [accessed 21 September 2023].

<sup>157</sup> Microsoft Bing, no date. [Bing Webmaster Guidelines](#). [accessed 16 June 2023].

- 13.37 A decision to deindex any URL that is found to contain illegal content would necessarily render *all* content found at the URL inaccessible to UK users via the service. In many cases, where the URL contained both legal and illegal content, this would result in UK users being denied access to lawful content. Such an approach would be likely to have detrimental commercial consequences for the controllers of URLs and databases and would engage the fundamental rights of interested persons, service providers and users alike to receive and impart information.
- 13.38 In addition, the degree of culpability of the URL controller may vary greatly from case to case. They may be a criminal, they may be a victim of hacking, or they may be responsible for content which is entirely lawful in their own country and is mostly targeted at their own country, but which is illegal content as the Act defines it. The prevalence of illegal content hosted by the interested person is likely to differ in each case. A recommendation to deindex in every case would preclude services from taking account of such nuances as they look to design their systems and processes for search moderation.
- 13.39 The Act tells us that certain offences are ‘priority offences’ which tends to suggest that illegal content of this nature should be treated more seriously.
- 13.40 On balance, therefore, we consider the impacts of a blanket deindexing recommendation would be difficult to justify where there may be less onerous means by which users may be protected from illegal content.
- 13.41 In cases where deindexing may not be appropriate or proportionate, providers should take other steps to protect users in relation to how search content is prioritised for users’ consumption. In practice, this means the downranking of URLs that host illegal content: if the material is (in principle) harder for users to find, it follows that it is less likely to cause them and others harm. In many cases, providers already deindex or downrank content that breaches providers’ terms of service or community guidelines.
- 13.42 To strike the right balance, we think it’s important for a provider to have discretion in designing its systems or processes for the purpose of moderating its search content. While a decision to deindex or downrank illegal content will in most cases contribute to the reduction of harm caused by the service (in respect of that content), the best approach to take will ultimately depend on the facts and circumstances.

### Factors relevant to deindexing or downranking

- 13.43 In respect of the first limb above regarding the making of illegal content judgments, we think the decision around whether to deindex or downrank (and, in the case of the latter, the decision around the extent of such downranking) should be taken having regard to (at least) factors which, as discussed above, are likely to be relevant to making a determination in any case, namely:
- a) the prevalence of illegal content hosted by the person responsible for the website or database concerned;
  - b) the interests of users in receiving any lawful material that would be affected; and
  - c) the severity of harmfulness of the content, including whether or not the content is priority illegal content.
- 13.44 Accordingly, this option would involve including wording to this effect in our draft Codes. Note that in Chapter 15 we make specific proposals which correspond to this proposed



measure for content detected using the ACM technologies we propose to recommend. This option does not affect those proposed measures.

- 13.45 The option would relate to the illegal content concerned, and so would normally impact at URL level rather than at domain level for general search services.

## Costs and risks

- 13.46 The costs of this option will vary by service. For small services with low risks which have few complaints, the costs could be low. Such services may have a process to assess all complaints of search results that include illegal content as they arise and deindex or downrank as appropriate, with the costs being low because they receive few complaints.
- 13.47 For services with significant risks of search results that include illegal content, the costs could be considerable as the volume of complaints about suspected illegal content could be high and the moderation systems and processes may need to be substantive.
- 13.48 However, we consider that the option outlined here is really the minimum that services would need to adopt in order to determine complaints and meet what is a fairly specific requirement in the safety duty, albeit subject to a proportionality threshold, to minimise the risk of individuals encountering search content which is illegal content. Incurring these costs is therefore necessary to meet the requirements of the Act.

## Rights impact

- 13.49 As set out in Chapter 12 in relation to U2U services' content moderation, search moderation is an area in which the steps taken by services as a consequence of the Act may have a significant impact on the rights of individuals and entities - in particular, to freedom of expression under Article 10 ECHR and to privacy under Article 8 of the European Convention on Human Rights ('ECHR').

## Freedom of expression

- 13.50 An interference with the right to freedom of expression must be prescribed by 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. Potential interference with users' freedom of expression arises where content is deindexed or downranked because the service considers it to be illegal content, particularly if that judgement is incorrect. However, as we have set out in Chapter 12, our starting point for the reason set out there is that Parliament has determined that services should take proportionate steps to protect UK users from illegal content. Of course, there is some risk of error in them doing this, but that risk is inherent in the scheme of the Act.
- 13.51 Services have incentives to limit the amount of content that is wrongly deindexed or deprioritised, to meet their users' expectations and to avoid the costs of dealing with appeals.
- 13.52 In addition, there could be a risk of a more general 'chilling effect' if users were to avoid use of services which have implemented a more effective moderation process as a result of this option. However, we do not consider that any such effect would be significant, given that many UK users already use services which have implemented moderation processes.

- 13.53 A greater interference would arise if the service, because of the Act, chose to adopt a publicly available statement which defined the content it would deindex or downrank more widely than is necessary to comply with 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 deindex or downrank content that is not or might not be illegal content, so long as they abide by the Act. Nothing in this option asks that services take steps against any content other than illegal content. Services have incentives to meet their users' expectations in this regard, too.
- 13.54 The duty for services to treat illegal content appropriately is a function of the Act, and not of this measure. This option is designed in a way that is not prescriptive about how illegal content is to be moderated, just that the provider's systems or processes are designed such that they deindex or downrank illegal content where they become aware of its presence on the service. It does not involve services taking any particular steps in relation to content of which they are not aware.
- 13.55 Impacts on freedom of expression could in principle arise in relation to the most highly protected forms of content, such as religious or political expression, and in relation to kinds of content that the Act seeks to protect, such as content of democratic importance and journalistic content. However, we consider there is unlikely to be a systematic effect on these kinds of content.
- 13.56 Where a service takes down content on the basis that it is illegal content, complaints procedures operated pursuant to section 32(2) of the Act allowing for the interested person to complain and for appropriate action to be taken in response may also mitigate the impact on their rights to freedom of expression.<sup>158</sup>
- 13.57 As set out above, the option relates to specific illegal content and as such, to comply with it, general search services should ordinarily implement downranking and deindexing measures at the URL level and not the domain level. We recognise that services are not required to do this and are permitted to take alternative approaches to complying with their duty about freedom of expression under section 33(2) of the Act (which requires them not to achieve particular outcomes but only to have particular regard to the importance of protecting users' right to freedom of expression within the law when deciding on and implementing safety measures and policies).

## Privacy

- 13.58 An interference with the right to privacy must be in accordance with the law and necessary in a democratic society in pursuit of a legitimate interest. Again, in order to be 'necessary', the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued. The content processed by search moderation functions is, by definition, either identified in a way that enables a general search service to pick it up, or is made available for publication by a vertical search service under a bilateral contract with the content provider. In either case, we do not consider that its review whether by an automated or a human search moderation function would amount to an interference with interested persons' rights to privacy under Article 8 ECHR. Any processing would need to be undertaken in compliance with

---

<sup>158</sup> See Chapter 16 (Reporting and complaints).

relevant data protection legislation (including, so far as the UK GDPR applies, rules about processing by third parties or international data transfers).

- 13.59 Interference with interested persons' or other individuals' privacy rights may also arise insofar as the option would lead to reporting to reporting bodies or other organisations in relation to illegal content. In particular, section 67 of the Act makes provision which (when brought into force) will require providers of regulated search services to report detected and unreported CSEA content to the Designated Reporting Body housed in the NCA (as further specified in the Act and to be specified in regulations made by the Secretary of State under section 68 of the Act). Providers may also have obligations to report CSEA content in other jurisdictions, or may have voluntary arrangements in place.
- 13.60 In part, any such interference results from the duties created by the Act or by existing legislation in other jurisdictions. In particular, where users or other individuals are correctly reported pursuant to the Act because they are suspected of committing the offence related to the CSEA content, any interference with their rights 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.
- 13.61 However, we have considered the extent to which the inclusion of this measure in our Codes of Practice as a recommended measure for the purpose of complying with providers' illegal content safety duties might give rise to additional interference.
- 13.62 Errors in content moderation decisions, whether made by automated technology or by humans, could result, in effect, in individuals being incorrectly reported to reporting bodies or other organisations, which would represent a potentially significant intrusion into their privacy. It is not possible to assess in detail the potential impact of incorrect reporting of users: the number of users affected would depend on what systems and processes the service implemented.
- 13.63 However, we do not consider it proportionate to expect all services, including very low risk, small and micro-businesses, to build in extra systems and processes to avoid accidental incorrect reporting to reporting bodies.<sup>159</sup> Reporting bodies have processes in place to triage and assess all reports received, ensuring that no action is taken in cases relating to obvious false positives. These processes are currently in place at NCMEC and will also be in place at the Designated Reporting Body in the NCA, to ensure that investigatory action is only taken in appropriate circumstances.

## Provisional conclusion

- 13.64 All search services must have proportionate systems and processes designed to minimise the risk of users encountering illegal content. They must also operate the service in a way that minimises the risk of individuals encountering search content that is illegal content. In practice, this means they must operate search moderation functions and should have systems or processes designed to deindex or downrank illegal content.
- 13.65 We acknowledge that there will be costs involved in search services operating these processes. We believe these costs are appropriate given the benefits they will create in reducing the risk of users encountering illegal content. Given that it is unlikely that a

---

<sup>159</sup> We consider what more might be needed for larger and riskier services below.

service could satisfy the requirements in the Act without implementing this measure, we regard the costs of this measure as primarily driven by the requirements of the Act, particularly given the considerable flexibility we have given to services as to how they implement the measure. If this measure is needed to meet the requirements of the Act, it must be a proportionate way to meet those requirements.

- 13.66 We considered whether it would be appropriate to recommend that a URL should always be deindexed where there were reasonable grounds to infer it contained illegal content. However, for the reasons set out above, we consider this would be disproportionate. We therefore propose that all search services should either deindex or, alternatively, downrank URLs where there are reasonable grounds to infer they contain illegal content. When deciding whether to deindex or downrank, they should have regard to at least the following factors:
- a) the prevalence of illegal content hosted by the interested person on the URL;
  - b) the interests of users in receiving any lawful material that would be affected;
  - c) and the severity of harmfulness of the content, including whether or not the content is priority illegal content.
- 13.67 As summarised at paragraph 13.3 above, rather than producing their own index, some search services (which we are calling 'downstream general search services') use the index produced by another large general search service rather than making their own. The level of control that a downstream general search service has over the index depends on the contract the provider has with the service they buy the index from. The nature of these contracts is not publicly known and is likely to differ from service to service.<sup>160</sup> Downstream general search services may not control the ranking of search content that might be accessed via their search engine and it may therefore not be possible for them to deindex illegal content directly.
- 13.68 However, we do not consider that different provision is needed for them as they can comply with their duties via their contract with the provider of their index. If complaints in relation to the downstream provider's service do not automatically pass to the upstream provider, the downstream service may need to make specific provision for them to be considered appropriately.
- 13.69 The duties in the Act apply to all search services, including downstream general search services and vertical search services. We therefore consider that this measure should apply to all search services.
- 13.70 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

## **Measure 2: Large general search services or multi-risk search services should set internal content policies having**

---

<sup>160</sup> In its advertising market study, the CMA said none of the contracts it had looked at allowed the downstream general search service to re-rank the search results it received from Google or Bing. Source: CMA, 2020. [Online platforms and digital advertising: Market study final report](#), Box 3.3 page 97 and paragraph 3.85.

## regard to at least the findings of their risk assessment and any evidence of emerging harms on their service

---

### Effectiveness

- 13.71 As set out in paragraph 13.15 above, general search services which are large may face significant challenges in terms of the volumes and diverse nature of the content they need to moderate. This could also be the case for any other search services that are multi-risk. This gives rise to questions about how such services should prioritise content for review, achieve consistency, quality and timeliness of decision-making, and plan their deployment of moderation resourcing so as to secure that users are appropriately protected. Accordingly, we have considered the case for including the following measure in our Codes:
- a) Large general search services or multi-risk search services should set internal content policies having regard to at least the findings of their risk assessment and any evidence of emerging harms on their service.
- 13.72 Search moderation relies on general rules, or ‘search moderation policies’, that in principle apply to all search content on a service. Policies are generally applied to individual URLs or domains, which for larger services is often done at scale.<sup>161</sup>
- 13.73 Search moderation policies may exist in two forms, external and internal. External policies are publicly available documents aimed at users of the service which provide an overview of a service’s rules about what content is allowed and what is not. These are often in the form of a publicly available statement. Internal policies are usually more detailed versions of external policies, and 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 designers of automated systems to assist in identifying potential breaches.<sup>162</sup> Search moderation policies can therefore help to secure more accurate and consistent decision making, particularly in organisations in which moderation is carried out by a large team.
- 13.74 In Chapter 12 we explained that there is 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 explained that there is a strong in-principle argument that where services are larger or higher risk and therefore need to moderate large volumes of diverse content, it is important that they have clear content moderation policies in order to ensure consistency, accuracy and timeliness of decision making. We consider that these arguments are likely to apply to general search services which are large and multi-risk in the same way that they apply to U2U services.
- 13.75 This suggests that for large or multi-risk search services, the existence of internal content policies is a pre-condition for being able to undertake effective search

---

<sup>161</sup> Google, no date. [Content policies for Google Search](#). [accessed 21 August 2023]; Cambridge Consultants, 2019. [Use of AI in Content Moderation](#). [accessed 3 August 2023]; Google, 2020. [Information quality & content moderation](#). [accessed 3 August 2023]; Ofcom, 2023. [Content moderation in user-to-user online services: An overview of processes and challenges](#).

<sup>162</sup> Khoury College at Northeastern University, no date. [Content Moderation Techniques](#). [accessed 3 August 2023]; Trust and Safety Professional Association, no date. [Policy Development](#). [accessed 3 August 2023]; Google, no date. [Content policies for Google Search](#). [accessed 17 August 2023].

moderation. We therefore consider that the option under consideration would deliver important benefits.

- 13.76 We also consider that there would be significant benefits in recommending that services have regard to at least risk assessments and evidence of emerging harms when setting their policies. Both of these data sources would provide evidence about the challenges search services' moderation functions face. It is reasonable to infer that such data would enable services to make higher quality decisions about what to put in their internal content moderation policies. This should improve the quality of these policies and by extension improve the performance of services' search moderation functions, thereby reducing the risk of harm to users.
- 13.77 For vertical search services, however, the benefits of this measure would likely be materially smaller. We have found no evidence so far of risks of illegal harms on vertical search services, and these search only for content provided by entities with whom they have a direct and ongoing contractual relationship. Given the lower risks, the volume and complexity of complaints about potentially illegal content which vertical search services receive is likely to be materially smaller than for general search services. Therefore the benefits of having internal content policies would be lower. We therefore would not consider it appropriate to apply this measure to large vertical search services just because they are large.

## Costs and risks

- 13.78 Services that do not currently have a search moderation policy covering all the matters set out above would incur the costs of developing it. This could involve legal, regulatory, as well as different ICT staff, and online safety/ harms experts. In some cases, services may use external experts which could increase costs. Agreeing new policies may also take up senior management's time which would add to the upfront costs. There would also be some small ongoing costs to ensure these policies remain up to date over time, as new risks emerge. We understand large general search services already have such policies in place, so in practice for such services this measure might only impose costs relating to ensuring their internal policies are sufficient to meet their duties under the Act.

## Rights impact

### Freedom of expression

- 13.79 The reasoning on the right to freedom of expression set out in relation to Measure 1 above applies equally in relation to this measure.
- 13.80 The option outlined here is designed in a way that does not tell services how to moderate illegal content, just that there are internal content policies outlining how to moderate it.
- 13.81 There is some risk that in writing their policies, services which align their terms and conditions with the definition of illegal content in the Act may over-generalise in a way which leads to over moderation. However, we consider that this risk arises equally if we were not to recommend this measure, since content moderators operating without any internal guidance may also over-generalise or be overly cautious.

- 13.82 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 and predictability of decisions, in a way which we think would also tend to protect users' rights to freedom of expression.

## Privacy

- 13.83 To the extent that, in setting content policies, services describe or define the content they are prohibiting in a way which involves reference to information in respect of which a user would have a reasonable expectation of privacy, or to personal data, users' rights in relation to these would be engaged.
- 13.84 However, that review (and the associated interference) is for the purpose of ensuring illegal content is taken down accurately for the purpose of the safety duty.
- 13.85 Where services are likely to be dealing with large volumes of content, the process of considering these matters in advance and preparing a policy would tend to improve internal scrutiny, and improve the consistency and predictability of decisions, in a way which we think would also tend to protect users' privacy and personal information rights.

## Provisional conclusions

- 13.86 For the reasons set out above, it is likely to be difficult for large general search services and any other search services that are multi-risk to carry out effective moderation without internal content policies. Given the importance of effective search moderation, we therefore consider that the benefits of a measure recommending that they put such policies in place would be substantial. Whilst we have not been able to quantify the costs, we understand that the largest search services already have content policies in place and are therefore in practice not likely to incur substantial new costs as a result of our proposal. Similarly, we believe that for small multi-risk services the costs of the proposal would be unlikely to be significant. On balance we therefore consider the proposed measure to be proportionate for these services.
- 13.87 Our analysis suggests that extending this measure to large vertical search services just because they are large would not confer significant benefits and that it is therefore unlikely to be proportionate. Similarly, the volume and complexity of moderation decisions for any search services that are neither multi-risk nor large appears likely to be much smaller than for large or multi-risk search services. Therefore the importance of having clear content policies would be much lower. Consequently, we do not propose extending the measure to any low-risk search services which are not large.

## Measure 3: Large general search services or multi-risk search services should set performance targets for their search moderation functions

---

- 13.88 We have considered the case for recommending the following measure:
- a) Large general search services or multi-risk search services should set performance targets for their search moderation functions and track whether they are meeting these. These should include targets for both how quickly search content that is illegal content is deindexed or downranked and for the accuracy of search moderation



decisions. When setting targets services should balance the need to deindex or downrank URLs containing illegal content swiftly against the need to make accurate moderation decisions. They should measure their performance against their targets.

## Effectiveness

- 13.89 In Chapter 12 we explained that many U2U services set performance targets for the operation of their content moderation functions and measure whether they are achieving these. We argued that setting performance targets and measuring whether they are achieving these is likely to deliver important benefits. We explained that where services are clear about the content moderation outcomes they are trying to achieve and measure whether they are achieving them, it stands to reason that they will be better able to plan how to configure their systems to meet these goals and better able to optimise the operation of these systems. We consider that these arguments are likely to apply to search moderation in the same way as for U2U moderation.
- 13.90 Consistent with the general approach described earlier in this chapter, we do not consider it appropriate to stipulate what the performance targets search services in scope of this measure should set. However, under the option we are looking at we *would* consider it appropriate that at a minimum these should include targets relating to the time within which they review or deindex/downrank URLs containing illegal content and targets relating to the accuracy of search moderation decisions. 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.
- 13.91 We consider that there would be important benefits to services setting both time based and quality/accuracy based targets for their search moderation teams and having regard to the desirability of striking a balance between timeliness and accuracy of decision making when setting their performance targets. Users are only protected if decisions are made in a timely way. Therefore there is a clear benefit to services having regard to the need for timely review of URLs containing potentially illegal content when setting their performance targets. At the same time, accuracy of decision making is also important and there is a strong case that a focus on speed of decision making should be balanced with a focus on accuracy. A disproportionate focus on speed of content removal could lead to pressure on systems which results in poorer quality decisions, which in turn could lead to a decrease in accuracy. As we set out in more detail in Chapter 12, a number of stakeholders have highlighted the importance of balancing speed and accuracy of both U2U and search services’ moderation decisions.

## Costs and risks

- 13.92 Services 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 the relevant information. There would also be ongoing costs. This would include data storage costs. More significantly, to assess the accuracy of search moderation decisions, services are likely to need to take a sample of those decisions and re-assessing them. There could therefore also be significant on-going costs from this measure.

- 13.93 We are not able to quantify these costs with any precision. They would depend in part on the complexity of the targets services set and the volume of content that was assessed.
- 13.94 There is a risk that setting performance targets could give rise to perverse incentives. For example, in principle there is a risk that unduly rigid targets could cause services to make sub-optimal decisions about which pieces of content to prioritise for review. However, we consider that the option considered is structured in such a way as to substantially mitigate this risk, given that we are allowing services flexibility for how to structure their targets and have explicitly set out that services should balance speed and accuracy of decision making.

## **Rights impact**

- 13.95 Our assessment of the rights impacts associated with Measure 1 also applies to this option in that moderating search results can infringe users' rights to free expression and, to the limited extent set out in relation to Measure 1, privacy. The risks to freedom of expression can be increased by the addition of performance targets in that a performance target relating to speed can cause moderators to try to take decisions quickly, increasing the risk of error and impacts on freedom of expression.
- 13.96 However, this option is designed to cause services to balance the need to take illegal content down swiftly with the need to make accurate moderation decisions. In particular, it does not specify a time within which decisions must be made, so the option should not put pressure on moderators to act so fast as to put users' rights to freedom of expression at risk.
- 13.97 The risks to privacy set out in relation to Measure 1, arising from the possibility that services may report detected illegal content to reporting authorities, are particularly acute where services are likely to be moderating content in large volumes. Whether automated technology is used, turnover of moderation staff, time pressures, seniority and experience of the person concerned can all affect the likelihood of error. We consider that the setting and monitoring of effective accuracy targets as a part of this option, also acts as a safeguard for users' rights to freedom of expression.

## **Provisional conclusions**

- 13.98 For general search services that need to make large volumes of moderation decisions, we consider there would be important benefits from setting performance targets for their search moderation functions and tracking whether they are met. As we explain above, we consider that services that follow this measure are more likely to operate effective search moderation systems. The evidence suggests that effective search moderation plays an important role in mitigating the risk of harm to users meaning the measure would have important benefits. As with Measure 2, these benefits will be greatest for general search services that are either large or multi-risk.
- 13.99 The costs of this measure are somewhat unclear. However, on balance, we consider that even in the context of this uncertainty, the benefits are likely to be sufficiently important to justify this proposal for large general search services and multi-risk general search services given the important role effective search moderation plays in protecting users from harm. That this proposed measure is proportionate is also consistent with Ofcom not proposing to be prescriptive on the details of the performance targets set or how

they are achieved. This leaves scope for services to tailor these targets according to the risks they identify and the specific operation of their services. This flexibility helps ensure that services can design performance targets and systems that are proportionate.

13.100 We are not proposing to extend this measure to large vertical search services just because they are large, nor to search services that are not multi-risk. This is because it is less clear the benefits are great enough given the lower risks associated with these services.

13.101 We therefore propose that:

- a) Large general search services or multi-risk search services should set performance targets for their search moderation functions and track whether they are meeting these. These should include targets for both how quickly URLs containing illegal content are deindexed or downranked and for the accuracy of search moderation decisions. When setting targets services should balance the need to deindex or downrank URLs containing illegal content swiftly against the need to make accurate moderation decisions. They should measure their performance against their targets.

13.102 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

## Measure 4: Large general search services or multi-risk search services should have and apply policies on prioritising content for review

---

13.103 Below we set out our analysis of the case for recommending the following measure in codes:

- a) Large general search services or multi-risk search services should have and apply policies on prioritising search content for review. In setting the policy, the provider should have regard to at least the following factors: (1) how frequently search requests for the search content are made ; the potential severity of the search content: including whether the content is suspected to be priority illegal content and the provider's risk assessment; (3) the likelihood that the search content is illegal content, including whether it has been reported by a trusted flagger.

### Effectiveness

13.104 Given the immense amount of content in the indexes they use, large or multi-risk general search services may need to deal with huge volumes of reports of URLs containing potentially illegal or otherwise harmful content. This means they will face difficult decisions about what search content to prioritise for review. The decisions they take about what to prioritise can have a material impact on the amount of harm a URL containing illegal content does to people. For example, if a general search service chooses to review a URL which was not appearing in many searches and which contained a small amount of relatively less egregious pieces of illegal content, before it reviewed a URL containing large volumes of extremely harmful illegal content that was appearing in lots of searches, this decision could result in significant harm to users.

13.105 We consider that where a service adopts a prioritisation framework which considers the factors listed above (as well as other factors they identify as relevant) this is likely to

result in high quality decisions about what search content to prioritise for review. Logically, we would expect this to result in a material reduction in harm to users compared to a counterfactual in which search services simply reviewed complaints in a chronological order, thereby delivering significant benefits.

13.106 We explain below why each of the prioritisation criteria covered by our option are important and relevant:

### How frequently search requests for the search content are made

13.107 The purpose of the Act is to make the use of regulated internet services safer for individuals in the United Kingdom.<sup>163</sup> Terms that are searched more often and by a greater number of users are likely to indicate a higher risk of harm to users. We therefore provisionally think services will achieve better outcomes for users if they have regard to search query frequency when prioritising search content.

13.108 We know that one large search service already considers the frequency with which certain queries are searched for by users when prioritising search content for review.<sup>164</sup>

13.109 However, we note that it is important to balance search query frequency alongside other factors, including those listed here, as prioritising only content which is searched for frequently may mean other very 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.

### Potential severity of the search content

13.110 Based on the evidence set out in paragraph 12.131 of the U2U content moderation chapter, we know that several U2U services already consider the severity (or egregiousness) of harm when prioritising content for review.<sup>165</sup> We expect the same to be true of search services. As set out in that section, some harms may be considered to have higher severity than others, such as those that have a degree of immediate direct harm compared to those that do not. All else being equal, addressing higher severity search content before lower severity content will minimise harm to users.

13.111 ‘Severity’ is also one of the three factors the UK Government used to determine its list of priority illegal offences and services should therefore consider these offences as high-severity.<sup>166</sup> However, services may determine that harms outside the list of priority illegal offences have a high-severity on their platform.

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

13.112 All else being equal, prioritising search content for review where the signals available to the service suggest that there is a high likelihood that it is illegal should increase the speed with which search content (such as URLs) containing illegal content are addressed, thereby reducing harm to users. Reasons to suspect that content is illegal can arise in a

---

<sup>163</sup> Section 1(1) of the Act.

<sup>164</sup> [CONFIDENTIAL ~~X~~].

<sup>165</sup> Ofcom, 2023. [Content moderation in user-to-user online services: An overview of processes and challenges](#). [accessed 18 September 2023].

<sup>166</sup> 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 2 August 2023].

number of different ways. Most obviously, users may complain about it. Their reports are likely to be the first and a very valuable way in which services may find out about illegal content, particularly for those services which are not making extensive use of proactive detection methodologies. However, in Chapter 12 we recognise that users are not always very good at correctly identifying breaches of U2U services' content policies. We consider the same is likely to be true of search services.

- 13.113 Dedicated reporting channels (DRCs), used by trusted flaggers<sup>167</sup> and Internet Referral Units<sup>168</sup>, are sometimes used by services to flag potentially illegal or violative content for review. In Chapter 16 we consider whether to recommend that search services establish a DRC for certain trusted flaggers relating to fraud.
- 13.114 Trusted flaggers can include internal teams, law enforcement, public sector organisations, civil society and private entities, and can offer particular expertise in notifying the presence of potentially illegal content on their website, which may result in higher quality flags or reports and potentially swifter removal of illegal content.<sup>169</sup>
- 13.115 Complaints are already commonly used to help prioritise content for review, and they can potentially flag illegal content that other search moderation functions may have missed. Where services have DRCs in place, the fact that a complaint comes from a trusted flagger is of obvious relevance in determining what priority to give it as, all other things being equal, such complaints are likely to be accurate and to reflect the trusted flagger's assessment of harm. They have significant potential to reduce harm to users.
- 13.116 In Chapter 15, we consider certain kinds of automated technology which are associated with a high likelihood that content they identify is illegal. Services may use other kinds of detection, whether human or automated, to identify content as suspected illegal content with varying degrees of certainty. The likelihood that the content is illegal is self-evidently relevant to whether further review is needed and how quickly it should take place.

## Costs and risks

- 13.117 The creation of a prioritisation framework would not in and of itself have an impact on the overall amount of search content general search services needed to review. However, there would be costs of designing and applying the prioritisation policy. The largely one-off costs of designing the prioritisation policy may take a small number of weeks of full-time work and involve legal, regulatory, as well as different ICT staff, and online safety/ harms experts, and agreeing the policy would likely need input from senior management. Applying that prioritisation policy could require system changes. For example, this might involve ensuring the virality of content is taken into account in 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. There would also probably be some smaller ongoing costs in ensuring

---

<sup>167</sup> Trusted flaggers are individuals, NGOs, government agencies, and other entities that have demonstrated accuracy and reliability in flagging content that violates a platform's Terms of Service. As a result, they often receive special flagging tools such as the ability to bulk flag content.

<sup>168</sup> Internet Referral Units are government-established entities responsible for flagging content to internet platforms that violates the platform's Terms of Service. Examples include the [EU Internet Referral Unit \(EU IRU\)](#) and the UK's [Counter Terrorism Internet Referral Unit \(CTIRU\)](#).

<sup>169</sup> European Commission, 2017. [Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms](#). [accessed 8 August 2023].

that the prioritisation policy is still reflected in system design, and in reviewing it when appropriate. These costs are mitigated by this option not specifying exactly how services should prioritise content, giving services some flexibility in what they do.

13.118 As the amount of content reviewed may not change it is not clear that establishing a framework for prioritising what to review having regard to the criteria set out here would impose other material ongoing content moderation costs on services compared to a counterfactual in which they simply reviewed complaints chronologically. Indeed, to the extent that general search services do not do this already, having a clear prioritisation framework may help them deploy their resources more efficiently.

## Rights impact

13.119 Our assessment of the rights impacts associated with having a search moderation function is set out above in relation to Measure 1. We do not consider that setting and applying a prioritisation policy would necessarily have any additional impacts on those rights. To the extent that it meant that harm would be a factor in services' decision making and that more users were better protected against harm, it is likely to result in a more proportionate approach to search moderation by the service, and therefore tend to safeguard users' rights.

## Provisional conclusions

13.120 For services that have a large quantity of potentially illegal content to review, there are likely to be significant benefits from prioritising that review in the way we propose, to reduce the harm from illegal content. While there are likely to be one-off costs of establishing a prioritisation system, we have not identified any large ongoing costs associated with the option. As the proposed measure does not specify exactly how services should prioritise search content, services have some flexibility to shape their approach to be proportionate to the risks that are on their service.

13.121 We do not consider that there would be significant benefits to extending this measure to large vertical search services just because they are large, nor to search services that are not multi-risk. We therefore propose to recommend the following measure in our codes:

- a) Large or multi-risk general search services should have and apply policies on prioritising search content for review. In setting the policy, the provider should have regard to at least the following factors: (1) how frequently search requests for the search content are made; (2) the potential severity of the search content: including whether the content is suspected to be priority illegal content and the provider's risk assessment; and (3) the likelihood that the search content is illegal content, including whether it has been reported by a trusted flagger.

13.122 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

## Measure 5: Large general search services or multi-risk search services should resource their search moderation functions sufficiently

---

13.123 We have considered the case for recommending the following measure in our Codes:

- a) Large general search services or multi-risk search services should resource their search moderation functions so as to give effect to their internal content policies and performance targets, having regard to at least: the propensity for external events to lead to a significant increase in demand for search moderation on the service; and the particular needs of its United Kingdom user base as identified in its risk assessment, in relation to languages.

13.124 We set out our analysis and findings below.

## Effectiveness

13.125 Where search moderation functions are adequately resourced one would expect this to enable them to review URLs containing potentially illegal content more quickly and make more accurate decisions as to whether to deindex or downrank them. We therefore consider that where search moderation functions are adequately resourced this will deliver significant and important benefits.

13.126 Setting objectives in relation to time and accuracy of a search moderation function as set out above would not protect users unless the service also set out to resource itself sufficiently, and deploy its resources effectively, so as to meet them. We therefore consider there would be significant benefits to large and multi-risk general search services resourcing their search moderation functions so as to meet these performance targets.

13.127 We do not at this stage think it would be beneficial for us to specify in detail how services should resource their search moderation functions. However, we do consider that there are factors to which services should have regard when deciding how to resource their search moderation function, and that considering these is likely to result in important benefits.

13.128 We explain below the factors we think services should consider and why each factor is important.

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

13.129 The evidence we have analysed suggests that to be effective, search services also need to build flexibility into their search moderation functions. In response to the 2022 Illegal Harms Call for Evidence, BSR stressed the importance of services ‘investing in the capability to scale-up/scale-down on short notice to respond to crisis events that can result in sudden spikes in illegal content.’<sup>170</sup> For example, search services may experience significant and sudden increases in search autocomplete complaints at times when there is significant public concern about a particular issue. Users may be at a heightened risk of encountering illegal content if services fail to take proportionate steps to plan for this.

13.130 Information obtained from platform risk assessments, tracking signals of emerging harm and other relevant sources of information, could be used to understand where and when such occurrences might happen.

---

<sup>170</sup> [BSR response](#) to 2022 Ofcom Call for Evidence: First phase of online safety regulation.



13.131 In instances where systems may need to deal with sudden harm events or significant and sudden increases in illegal search content, redeploying resource may draw resource away from another part of the system. Services which have contingency plans in place to ensure that illegal content across the system is dealt with expeditiously are more likely to protect users effectively. Hence it would be beneficial if general search services considered the potential for significant and sudden increases in problematic and potentially illegal content when determining how to resource their search moderation functions.

### The particular needs of its United Kingdom user base as identified in its risk assessment, in relation to languages

13.132 In paragraphs 12.158-12.164 of the U2U content moderation chapter, we set out evidence that U2U services moderate multilingual content, and the importance of services being able to deal with different languages and understand cultural context.<sup>171</sup> Reflective of our analysis and conclusions there, as well as the fact that we know users in the UK use search services in multiple languages, deploying appropriate language resource and expertise (such as moderators with language expertise or automated systems that work in the required language) would enable services to identify, review and moderate search content that is suspected illegal content.<sup>172</sup>

13.133 The language expertise required to deal with the risk of harm in a particular language will likely differ from service to service based on a number of factors, including user base and whether the service is a general search service or a vertical search service. For this reason, we feel Codes should not be prescriptive around what exact language expertise and resource is required on any service. However, where services deploy appropriate language resource and expertise (such as content moderators with language expertise or automated systems that work in the required language) in a way that enables them identify, review and remove content in accordance with Measure 1 above, users would be materially better protected than in a counterfactual where a search moderation function did not have the appropriate language expertise available to it. This would deliver important benefits. It should be noted that the Online Safety Act is concerned with protecting users of services in the UK, so any recommendation would be in relation to languages used or viewed by users of services in the UK.

### Costs and risks

13.134 The costs of resourcing services' search moderation systems and processes to give effect to internal content policies and meet performance targets is likely to be substantial and ongoing. It will tend to be higher, the higher the volume of webpages included in the index, which we understand tends to be higher for larger services.<sup>173</sup>

---

<sup>171</sup> 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.' Source: Carnegie UK, 2021. [Ad hoc advice to the United Nations Special Rapporteur on Minority Issues](#). [accessed 3 August 2023].

<sup>172</sup> Vox, 2015. [In which language do you Google? Tracking 135 languages in 9 cities since 2004](#). [accessed 17 August 2023].

<sup>173</sup> 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 21 September 2023].

13.135 The type of detection and review processes are likely to influence the magnitude of costs:

- E.g. automating moderation processes (e.g. machine learning solutions for AI) require both one-off infrastructure investment, and different ICT professionals' time. Additionally, system updates, and licensing costs can be expensive and add to ongoing costs.
- If search moderation involves human moderators, resourcing costs will primarily depend on how many moderators are needed.
- Some services may require a separate review process for more complex illegal content cases, which may also require legal input.<sup>174</sup>

13.136 While these costs will be significant for some services, the option outlined does not include specific outcome targets for services. It would therefore be for services to determine how much they need to do, including the duties on them in the Act.

## Rights impact

13.137 Our assessment of the rights impacts associated with having a search moderation function is set out above in relation to Measure 1 and our assessment of the implications of having performance targets is set out above in relation to Measure 3. We do not consider resourcing the function appropriately would have any additional impacts on those rights.

## Provisional Conclusions

13.138 In view of the analysis above, we propose to recommend the following measure in our codes:

13.139 Large general search services or multi-risk search services should resource their search moderation functions so as to give effect to their internal content policies and performance targets, having regard to at least: the propensity for external events to lead to a significant increase in demand for search moderation on the service; and the particular needs of its United Kingdom user base as identified in its risk assessment, in relation to languages.

13.140 Our analysis suggests that this measure could impose significant costs on services. However, for the reasons we explain above, we consider that if search moderation functions are not adequately resourced having regard to these factors this could significantly reduce their effectiveness. Given the importance of effective search moderation, this could give rise to very significant harm. While our proposed measure requires services to resource their search moderation functions to give effect to their performance targets, we do not propose to specify precisely what those performance targets are, which gives services some flexibility in precisely what they do. We therefore provisionally consider that this recommendation would be proportionate.

13.141 We are not at this point proposing extending the measure to large vertical search services just because they are large, nor to search services that are not multi-risk. The

---

<sup>174</sup> "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 Ofcom Call for Evidence: First phase of online safety regulation](#). [accessed 21 September 2023].

lower the risks associated with a search service the less potentially illegal content it is likely to need to address. Therefore, the benefits of applying this measure to large vertical search services just because they are large and multi-risk search services would be materially smaller than for large general or multi-risk search services. Given the significant costs of the measure, it is not clear it would be proportionate to recommend this measure more widely. Moreover, this measure is predicated on services having the internal content policies of Measure 2 above and the performance targets we propose in Measure 3, so it makes sense for this measure to apply to the same set of services as those proposed measures are recommended for.

13.142 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

## Measure 6: Large general search services or multi-risk search services should train people involved in search moderation and provide materials

---

13.143 As set out in relation to Measure 1, in order to comply with the Act, a service considering suspected illegal content should either make an illegal content judgment in relation to it, or, if it is satisfied that its PAS for the 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. It follows that the moderators carrying out this work need to know how to do whichever of those two things the service has chosen to do.

13.144 For small, low risk services which moderate little content, it may be possible to do this without training or written materials. But for services which are subject to Measure 2, we consider it very unlikely that it would be possible for moderators to give effect to search moderation policies without training and additional materials (such as: definitions and explanations around specific parts of the search moderation policy, enforcement guidelines, examples, and visuals of the tool or interface moderation staff will use to carry out their job). The extent of illegal content that larger and riskier services may face, as set out in paragraph 13.15 above, is far greater.

### Option(s) and Effectiveness

13.145 In this section, we are considering an option of recommending that services which have search moderation policies should ensure that people working in its content moderation process receive training and materials that enable them to moderate content in accordance with Measures 1 and 2.

13.146 There is limited evidence on how search services train staff (including contractors, etc.) involved in content moderation. We know that some larger services train their moderators and other relevant members of staff to identify and action illegal content, as well as providing supporting materials to help them do so. Nevertheless, we believe it would be similar to how this is carried out by user-to-user services (see Measure 6 from paragraph 12.173 of Chapter 12).

13.147 In its response to the 2022 Illegal Harms Call for Evidence, [CONFIDENTIAL X].<sup>175</sup>

---

<sup>175</sup> [CONFIDENTIAL X].

- 13.148 As explored in the U2U content moderation chapter, a number of respondents to the 2022 Illegal Harms Call for Evidence, particularly civil society organisations, as well as broader academic and civil society literature and research, stress the importance of training moderation staff, as well as the importance of providing staff with materials that support them in minimising the risk of users encountering illegal content (see paragraph 12.175-12.213 for full analysis).
- 13.149 Based on the information above, as well as analysis carried out in the user-to-user content moderation chapter, we consider that training staff involved in moderation, as well as providing them with relevant materials, is beneficial for identifying and minimising the risk of users encountering illegal content, especially when compared to not training staff. Staff that have been trained on how to identify and action content in accordance with Measure 1 above are more likely to be equipped with the knowledge and skills to identify when action needs to be taken against search content, when compared to those who are untrained.
- 13.150 We also think that staff involved in moderation who are trained regularly will have up-to-date knowledge of content moderation policies, as well as on the systems they are using to carry out their job.
- 13.151 There is no set best practice on how often training or supporting materials should be refreshed, and it may depend on a number of factors, including a person's role and performance. However, if moderators are trained on any major changes to policies or processes relating to content moderation, and provided with new or updated supporting materials, they are more likely to be able to give effect to them accurately and consistently.
- 13.152 We therefore provisionally consider that users would be better protected from harm if we recommend that a search service which has a moderation policy should ensure that people working in its moderation process receive training and materials that enable them to moderate in accordance with Measures 1 and 2.
- 13.153 However, we do not currently consider that the option we have outlined here should include voluntary content moderators for the same reasons discussed in the U2U content moderation chapter (see 12.189). We are unaware of any search services that employ volunteer moderators and therefore we do not envisage this will currently impact any service.
- 13.154 We do not consider that it would be appropriate to specify in Codes how often materials should be revised or training should be redelivered. A service which failed to refresh training and materials following any major changes to policies or processes relating to content moderation that is to do with suspected illegal content would not be enabling its moderators to moderate content in accordance with Measures 1 and 2, in particular, Measure 2.
- 13.155 As set out above, we consider that generally speaking services are best placed at present to determine what is appropriate for their services in terms of the detail of their training and materials. However, services which do not have regard to certain factors are unlikely to protect users properly. We therefore consider below whether to specify in Codes that in preparing and delivering search moderation training and materials, services should have regard, at least, to matters we specify in Codes.

## Possible factors to consider in the training of staff involved in content moderation and supporting materials

- 13.156 We have provisionally identified a number of matters which we consider likely to be relevant to services' decisions about how they should train their search moderation functions.
- 13.157 **Risk assessment and information pertaining to the tracking of signals of emerging harm** - A service's risk assessment will be one of the key sources of information telling a service what risk of search content that is illegal content they have on their platform and would form the basis for internal content policies (see Measure 2). As moderators should be focused on enforcing the internal content policies, training should also be informed by the most recent illegal content risk assessment. In Chapter 8, we are also consulting on a proposed recommendation that services should track signals of emerging harm. If, following consultation, we remain of the view we should recommend this, this information would be one of the key sources of information about how illegal content manifests and it is therefore crucial services use this to inform their content moderation training and supporting materials.
- 13.158 **Remedying gaps in moderation staff's understanding of specific harms** – There may be instances where staff do not have the appropriate 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 identifying and actioning search content that is illegal content due to the unique, complex, novel or serious nature of a given harm. For example, although some CSAM can be easily identified as illegal content, there are many exceptions to this. For example, it can be difficult for content moderators to determine whether an image depicts a person who is under or over 18. Specific training should be provided to those involved in content moderation of such content. If training and materials are given to moderators where a service has identified a gap in moderators' understanding of a specific harm, and where they deem there to be a specific risk, this should improve outcomes for users.
- 13.159 **Staff welfare** - As set out in Chapter 12, we do not currently have evidence to suggest that staff welfare matters affect user safety, and do not propose to make recommendations on this in our first Codes.

## Costs and risks

- 13.160 The main factors driving the cost of the training would be the number of staff to be trained and the duration of the training. Our analysis of this is the same as that for U2U services content moderation chapter (see paragraphs 12.200 to 12.204 for this analysis). In summary, we estimate that the costs of providing training for one new content moderator could be between £2,500 and £15,000, and for a new software engineer between £3,500 and £21,000.
- 13.161 As the number of moderators that need training is likely to depend on the volume of content that needs to be assessed, the costs of this measure are likely to scale with the benefits.

## Rights impacts

### Freedom of expression

- 13.162 As several respondents to the 2022 Illegal Harms Call for Evidence noted, training enables those involved in content moderation to make better decisions. [**CONFIDENTIAL** ✕]. Training also enables staff involved in moderation to have a better understanding of borderline content, i.e. content where it can be difficult to determine whether it is legal or illegal. All things being equal, better training should safeguard users' rights to freedom of expression.

### Privacy

- 13.163 Services would need to comply with privacy and data protection laws in relation to any items of content they use in their training and other materials.
- 13.164 We consider that the training of moderators would be a further safeguard for users' privacy, against the possibility that services may incorrectly report detected illegal content to reporting authorities.

## Provisional conclusion

- 13.165 As set out above, this option is linked to and would be effective for those services which have search moderation policies in compliance with Measure 2. It follows that it should only be considered for those services – i.e. large general search services or multi-risk search services.
- 13.166 We recognise that the additional costs may be significant for some services. However, we also consider that the benefits of this measure are likely to be high. This is because moderator training is important in effectively implementing a service's search moderation policies to reduce harm and comply with its online safety duties. Well-trained and prepared moderators are much more likely to be able to identify content in accordance with Measure 1 and the service's content standards under Measure 2, and to apply the correct treatment to it, materially reducing the harms that result from that. As such, this measure is likely to be proportionate for services which identify significant risks to users.
- 13.167 We consider this to be the case even for smaller services. Training costs are likely to depend primarily on the number of people that need to be trained. Everything else being equal, smaller services are likely to have smaller volumes of content (or smaller volumes of complaints) to review, and fewer moderators as a result. This means the costs for smaller services will be correspondingly lower than for large services.
- 13.168 For these reasons, our provisional view is that it is proportionate to recommend this measure to large general search services or multi-risk search services.
- 13.169 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.
- 13.170 The full text of our proposed measure, covering each of the factors outlined in our discussion above, can be found in our proposed code of practice, Annex 8, Recommendation 4F.

# 14. Automated Content Moderation (U2U)

## What is this chapter about?

In our Content Moderation (U2U) chapter, we explained our proposals in relation to the measures services should take to set up their content moderation systems in a manner consistent with the safety duties. We explained that services use automated tools, often in tandem with human oversight, to make content moderation processes more effective at identifying and removing illegal and violative content. As these tools allow services to surface large volumes of harmful content at pace, they are critical to many services' attempts to reduce harm. This chapter focuses in detail on automated content moderation tools, and what automated tools our Codes should recommend U2U services use.

## What are we proposing?

We are making the following proposals for certain U2U services:

**We propose to recommend that 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 should apply to the following services:**

- large services which are at medium or high risk of image-based CSAM in their risk assessment;
- other services which are at high risk of image-based CSAM in their risk assessment and have more than 700,000 monthly UK users;
- 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.

**We propose to recommend that certain types of service should use an automated technique known as URL detection to analyse relevant content to assess whether it consists of or includes a CSAM URL, and should take appropriate measures to swiftly take down those URLs detected. This measure 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.

**Articles for use in frauds (standard keyword detection): the following types of service should put in place standard keyword detection technology to identify content that is likely to amount to a priority offence concerning articles for use in frauds (such as content which offers to supply individuals' stolen personal or financial credentials), and consider detected content in accordance with their internal content moderation policy. This measure would apply to the following services:**

- large services which are at medium or high risk of fraud in their risk assessment.

These proposals only apply in relation to content communicated **publicly** on U2U services, where it is technically feasible to implement them. Consistent with the restrictions in the Act, they do not



apply to private communications or end-to-end encrypted communications. In Annex 9 to this consultation, we have set out draft guidance which is intended to assist services in deciding whether content has been communicated “publicly” or “privately” for this purpose.

## Why are we proposing this?

### CSAM

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.<sup>176</sup> While our proposals would 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 provisionally consider that, even where they are very small, it could be justified to recommend that services which are high risk to deploy these technologies. However, we are proposing to set user-number thresholds below which services would 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 proposed 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.

We propose setting a lower threshold for file-storage and file-sharing services because there is evidence to suggest that this kind of service plays a particularly significant role in the circulation of CSAM. Further, file-storage and file-sharing services typically reach a lower number of users than some other kinds of service. We therefore consider it appropriate to set a lower threshold for file-storage and file-sharing services to ensure they are not out of scope of the measure despite the significant role they play in the circulation of CSAM.

### Fraud

Fraud is the most commonly experienced illegal harm, and it can cause significant financial and psychological harm. Our research shows that some services are being used by fraudsters to supply, or offer to supply, articles for use in frauds (including stolen personal and financial credentials). Not only is this a priority offence, but it can facilitate other priority illegal fraud offences. Our research also indicates that, when discussing such articles, very specific keywords tend to be used, and that – particularly when combined - these are unlikely to be used in any legitimate context.

Our provisional view is that standard keyword detection technology would be an effective means to proactively identify content likely to amount to an offence concerning articles for use in frauds. Such content would then be considered by services in accordance with their content moderation policies. Whilst our proposal would impose significant costs on some services, we consider this justified given the very serious nature of the harm it addresses. To ensure the costs are proportionate, we propose targeting this measure at large services with a medium or high risk of fraud.

---

<sup>176</sup> Though we note there are limits to what they can achieve, in the context of eradicating CSAM online.

The automated tools we propose including in this version of our Codes are well-established and have been used for years by many of the larger services. In practice, there is a range of significantly more sophisticated automated tools which services use to detect harmful content, including natural language processing and the use of machine learning to identify new previously undetected harmful content. Such tools play an important role and we do not wish to discourage their use; indeed we are supportive of industry efforts to develop and refine them. However, we do not have sufficient evidence on their costs and efficacy at this stage to justify including provisions relating to their use in the first version of our Codes of Practice.

### What input do we want from stakeholders?

- Do you agree with our proposals? Do you have any views on our three proposals, i.e. CSAM hash matching, CSAM URL detection and fraud keyword detection? Please provide the underlying arguments and evidence that support your views.
- Do you have any comments on the draft guidance set out in Annex 9 regarding whether content is communicated ‘publicly’ or ‘privately’?

Do you have any relevant evidence on:

- The accuracy of perceptual hash matching and the costs of applying CSAM hash matching to smaller services;
- The ability of services in scope of the CSAM hash matching measure to access hash databases/services, with respect to access criteria or requirements set by database and/or hash matching service providers;
- The costs of applying our CSAM URL detection measure to smaller services, and the effectiveness of fuzzy matching<sup>177</sup> for CSAM URL detection;
- The costs of applying our articles for use in frauds (standard keyword detection) measure, including for smaller services; and
- An effective application of hash matching and/or URL detection for terrorism content, including how such measures could address concerns around ‘context’ and freedom of expression, and any information you have on the costs and efficacy of applying hash matching and URL detection for terrorism content to a range of services.

## Introduction

---

- 14.1 In Chapter 12 we discuss our approach to content moderation for U2U services. Here we focus in detail on what we describe as automated content moderation (ACM).
- 14.2 As the amount of user-generated content on platforms continues to increase rapidly, it is not possible for many services to identify and remove illegal content using traditional human-led moderation approaches at the speed and scale necessary. ACM tools can support content moderation by automating the review of content – either when it is uploaded or once it is on the service – including through comparing each piece of content against a database of known illegal or other harmful content. A piece of content that matches existing content in the database can then be flagged for further review or automatically removed,

---

<sup>177</sup> Fuzzy matching can allow a match between U2U content and a URL list, despite the text not being exactly the same.

depending on the setup of the tool. These are important tools that we know services use, often in tandem with human moderators, to make content moderation processes more effective at identifying and removing illegal or otherwise harmful content.

- 14.3 Depending on what harm these tools are being applied to, and in what way, their accuracy, effectiveness and degree of bias can vary. They can therefore have a significant impact on user rights, in particular freedom of expression and privacy. They can also incur significant costs, varying depending on the nature and complexity of the technology and how it is applied. As such, we have assessed each of the tools we are proposing in detail.
- 14.4 The main types of ACM technologies considered in this chapter are as follows:
- a) **Hash matching** is a process for detecting when users attempt to upload content which has previously been identified as being illegal or otherwise violative. It allows services to prevent the re-upload of illegal content. It involves matching a hash of a unique piece of known illegal content stored in a database with user-generated content. Hashing is an umbrella term for techniques to create fingerprints of 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. Hash matching can be used to prevent the upload, download, viewing or sharing of illegal or harmful content.
  - b) **Uniform Resource Locator (URL) detection** is a process by which URLs (i.e., individual webpage addresses) known to host illegal content are matched against user-generated content. Whilst hash matching enables a service to detect illegal or harmful files present on that same service, URL detection allows a service to detect previously identified links to illegal or harmful content on other services. In this sense, it is complementary to hash matching.
  - c) **Standard keyword detection** involves the use of words and/or phrases that are indicative of a particular harm or offence (e.g. 'Fullz' for stolen credentials fraud). The words and phrases can be used by services to detect illegal content and content violative of their terms of service.
- 14.5 For each of the above applications, once a match of some form is established, the content can either undergo human review or be removed automatically.
- 14.6 In addition, some services also use machine learning (ML) to detect previously unidentified illegal content, sometimes in conjunction with the more straightforward technologies listed above.
- 14.7 Our proposed recommendations focus on technologies where we currently have sufficient evidence and understanding to give us confidence that they will be effective and recommending their use is proportionate in the circumstances. In particular, we have considered evidence of how these technologies can be applied to specific harms. We are aware that many services make use of technologies that utilise alternate processes, such as Artificial Intelligence (AI) and ML. However, given limited evidence available at this stage, we are not proposing to include recommendations for these technologies in this first version of the Codes.
- 14.8 This chapter outlines various ACM technologies and each of our specific policy recommendations. We begin by setting out our overall approach to assessing the ACM measures before assessing each measure in depth. For each measure, we provide a technical description of the technology and an overview of the harm that the measure seeks to address. We then set out a measure (the design of which is explained in Annex 15) and our

assessment of the measure's accuracy, effectiveness, lack of bias, costs, risks and potential impact on rights. The chapter ends outlining some initial thoughts on the use of AI to detect previously unidentified illegal content, and on cumulative risk scoring systems.

- 14.9 The Code measures that we are proposing are recommended for the purpose of compliance with providers' safety duties under sections 10(2) and (3) of the Act. The measures each relate to kinds of priority illegal content or to the risk of the service being used for the commission or facilitation of a priority offence. The measures concerning hash matching for CSAM and the detection of CSAM URLs would form part of the CSEA Code of practice.

## Our approach to looking at the measures

- 14.10 Our approach to ACM measures is distinct from Chapter 12 on content moderation for U2U services and goes beyond providing guidance on the factors services should have regard to when designing and deploying their content moderation systems and processes. We believe this approach is warranted for the following reasons:
- a) **The additional constraints in the Act concerning measures that qualify as proactive technology** means that there is a requirement for us to assess in greater detail their accuracy, effectiveness and lack of bias.
  - b) **The nature of the harms identified** we consider to be particularly serious and requiring of complex measures. Given this, we consider that these measures require a higher degree of specificity to support services in adopting them. This will also aid us in ascertaining whether services have adopted them.
  - c) **Our current level of understanding and evidence** is strongest in the selection of measures that we have proposed. While this demonstrates that we have greater confidence in the measures as proportionate steps to mitigate harm, it also explains why we have considered but not adopted ACM measures across different harms, and in different circumstances, where we currently do not have the appropriate level of evidence to gauge proportionality.
- 14.11 Chapter 11 sets out our approach to developing recommended measures. In view of these considerations, our proposals recommend the use of a kind of technology but do not recommend specific technologies or use of specific inputs (such as a hash database or URL list provided by a specified third party). We aim instead to set out our proposals in sufficient detail to ensure that they are effective and that services are readily able to adopt them (consistent with the principle that measures should be sufficiently clear, and at a sufficiently detailed level, that providers understand what those measures entail in practice).<sup>178</sup> This would ensure that services can act in accordance with our recommendations using any appropriate technology or input.
- 14.12 We are not proposing to recommend some measures which may be effective in reducing risks of harm. This is principally due to currently limited evidence regarding the accuracy, effectiveness and lack of bias of the technologies that the measures refer to. We recognise that some of these measures may be proportionate for certain services to take, and welcome further innovation and investment in safety technologies to support ACM. We plan to consider further ACM measures for future versions of our Codes.

---

<sup>178</sup> See paragraph 2 of Schedule 4 to the Act.

## Proactive technology

- 14.13 Section 231 of the Act makes clear that automated content moderation technologies can fall within the definition of what the Act refers to as ‘proactive technology’. This is important because paragraph 13 of Schedule 4 to the Act contains a number of constraints on Ofcom’s ability to recommend the use of ‘proactive technology’ in Codes of Practice.
- 14.14 These include that:
- a) Ofcom may not recommend in a Code of Practice the use of the technology to analyse user-generated content communicated “privately”, or metadata relating to user-generated content communicated “privately”.<sup>179</sup>
  - b) When deciding whether to include a proactive technology measure in a Code of Practice, we must have regard to the degree of accuracy, effectiveness and lack of bias achieved by the technology.
  - c) A proactive technology measure may be applied to services of a particular kind or size only if we are satisfied that the use of the technology in question by such services would be proportionate to the risk of harm that the measure is designed to safeguard against (taking into account, in particular, the risk profile relating to such services).
- 14.15 Our proposals in this chapter take account of these constraints.
- 14.16 Our proposed recommendations in this chapter will also only apply where it is technically feasible for a service 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 would not apply to services 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. For example, we acknowledge that end-to-end encrypted services are currently unable to analyse user-generated content in the ways set out in our proposals.
- 14.17 The Act imposes other requirements on services in connection with their use of proactive technologies. Chapter 18 discusses service providers’ duty to include provisions in the terms of service giving information about any proactive technology used for the purpose of compliance with the illegal content safety duties.

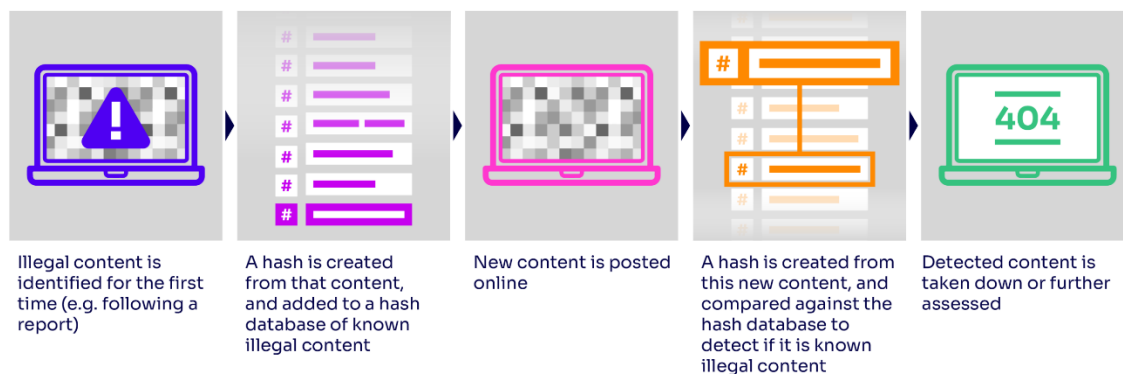
---

<sup>179</sup> In Annex 9, we have set out draft guidance which is intended to assist services in deciding whether content is communicated ‘publicly’ or ‘privately’ for this purpose.

# Hash matching

## Introduction

Figure 14.1: Overview of hash matching



Source: Ofcom

- 14.18 Hash matching is a form of automated content moderation which can be used to detect illegal content. It operates by comparing a digital fingerprint of a file, known as a ‘hash’, to a hash created previously from another file. As such, it is only capable of detecting matches to ‘known’ illegal content which has already been identified and hashed.
- 14.19 A hash is created from a file using an algorithm known as a ‘hash function’. For images and videos, hash functions fall into two types: cryptographic hash functions use the so-called ‘avalanche effect’ to create very different hashes where the input files differ only a little, while perceptual hash functions aim to create very similar hashes from very similar files.
- 14.20 These types of hash function support different kinds of hash matching:
- **Cryptographic** hash matching determines whether a given file is identical to a hashed file. If the hashes match, the two files can be taken to be identical. Cryptographic hash matching is highly accurate in detecting matches between two identical pieces of content. However, it will not detect a match where a file has been modified from the file originally hashed, which reduces its effectiveness in detecting illegal content and renders it vulnerable to deliberate evasion.
  - **Perceptual** hash matching determines whether a given file is likely to be perceived as similar to a hashed file. It does this by comparing the similarity between the hashes of the files, assessed using a ‘distance metric’.<sup>180</sup> A threshold is set to determine when there is sufficient similarity between the hashes for the files to be considered a (near) match – that is, perceptually similar to each other. This allows modifications from an original file to be detected. Perceptual hash matching can

<sup>180</sup> The similarity between any two hashes is defined through a distance metric (e.g. the Euclidean distance), and different perceptual hash functions may require the use of different distance metrics. The goal is to approximate the level of similarity between input files perceived by humans through the distance between / similarity of the perceptual hashes. Source: Ofcom, 2022. [Overview of Perceptual Hashing Technology](#).

therefore be more effective in detecting illegal content, but with an increased probability that content may be incorrectly detected as a match for illegal content.

- 14.21 The effectiveness of both kinds of hash matching in detecting illegal content depends on the database of hashes of known illegal content which is used to compare content against. The more illegal content included in the hash database, the more effective hash matching can be in detecting illegal content.
- 14.22 Equally, the accuracy of both kinds of hash matching depends on whether content has been correctly added to the hash database. If content in the database was mis-classified as illegal when it was added, this could lead to content on a platform detected as a match being wrongly treated as being illegal. Ensuring the integrity of a hash database for illegal content requires governance arrangements to ensure that decisions about whether content is illegal (which can involve the making of difficult legal or factual judgements) are properly made and the database is secured against unauthorised access.
- 14.23 It should be noted that hash databases for illegal content may deliberately include only certain kinds of illegal content (for instance, only terrorism content that meets specified policy criteria), or be operated by reference to the criminal law of another jurisdiction. The operators of hash databases may also deliberately choose to include content that is not illegal but is considered to be harmful (or is considered appropriate to include for other reasons).
- 14.24 We are aware of a number of accessible options for services seeking to implement hash matching technology or to obtain hashes from external sources, such as from national and international NGOs. Services can take different approaches depending on their capacity, resources and whether they can access third party provider hash databases. In some cases, services will use a combination of approaches to ensure they are comparing user-generated content to a very broad range of hashes and using multiple different hash functions. In other cases, services may not do any hash matching themselves but will procure a third party to provide the database, the hashing function and conduct the hash matching process.
- 14.25 The performance of hash matching will also depend on the hash function used and, in the case of perceptual hash matching, the distance metric and threshold used to identify when two hashes are sufficiently similar to constitute a match.
- 14.26 The concepts of **false positives** and **false negatives** are relevant here. In the context of detecting matches for illegal content, a false positive is a case where the technology has incorrectly identified content as a match for illegal content, and a false negative is a case where the technology has not detected content as a match when it is in fact an exact or near match for known illegal content.
- 14.27 False positives and false negatives can be reflected in measures of statistical accuracy, including:
- **precision** – which refers to the percentage of content detected as positive that is in fact positive; and
  - **recall** – which refers to the percentage of all content that is in fact positive which is detected as positive.
- 14.28 There is a trade-off between precision and recall. A perceptual hash matching system which seeks to find as much illegal content as possible (maximising recall) may result in an increased level of false positives (lowering precision). Conversely, a system which seeks to



minimise false positives (maximising precision) will detect less illegal content (lowering recall).

- 14.29 In making this trade-off, the prevalence of illegal content is important. If illegal content makes up only an extremely low percentage of all content, a high proportion of detected content could be false positive results even with a technology which appears to have a low false positive rate (because there are relatively few items of illegal content to be found, and many opportunities for the technology to wrongly identify other content as a match).
- 14.30 The level of false positives (including any cases arising from content being incorrectly included in a hash database) determines the potential impacts on users' freedom of expression and privacy. These impacts can be addressed by steps such as further review of detected content by human moderators before action is taken in response to it, or the operation of a complaints procedure which enables users to complain if they believe their content has wrongly been identified as illegal content.

## Hash matching for child sexual abuse material (CSAM)

---

- 14.31 This section explains our consideration of the case for recommending certain services use hash-matching technology effectively to detect known CSAM in the form of images or videos, which is (or would be) communicated publicly by means of the service, and swiftly take it down.
- 14.32 '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.<sup>181</sup>
- 14.33 Indecent or prohibited images include still and animated images, and videos, and can include photographs, pseudo-photographs, and non-photographic images such as drawings, which depict sexual activity with a child or are otherwise indecent. Other material could be in any form, including images, video, written or audio content. However, subsequent references to CSAM in this section are to CSAM in the form of images and videos.

## Harms that the measure seeks to address

- 14.34 The prevalence of CSAM online and the extent of its dissemination is difficult to quantify. However, the Internet Watch Foundation (IWF) has described it as growing 'exponentially'.<sup>182</sup> Some Non-Governmental Organisations (NGOs) gain insight into the amount of CSAM online by operating online reporting systems. This allows users and online services to report instances of CSAM to an NGO who can pass on reports to law enforcement. Some NGOs also run hash matching databases and technologies, and collect metadata<sup>183</sup> This is combined into usable insights to examine online trends in CSAM

---

<sup>181</sup> For further detail, see Chapter 5. Child sexual abuse and exploitation (CSEA): Offences relating to child sexual abuse material (CSAM) of the draft Illegal Content Judgements Guidance, published as Annex 10 of this consultation.

<sup>182</sup> IWF, 2021. [accessed 7 June 2023].

<sup>183</sup> For example: IWF, 2023. [Image Hash List](#). [accessed 26 May 2023]; National Centre for Missing and Exploited Children (NCMEC), 2022. [CyberTipline 2022 Report](#). [accessed 26 May 2023].

distribution and circulation. As outlined in the Register of Risk,<sup>184</sup> the National Center for Missing and Exploited Children (NCMEC) in the US received over 32 million reports to its CyberTipline in 2022, up from 29.4 million in 2021 and 21.8 million in 2020.<sup>185</sup> Over 99% of these reports depicted suspected CSAM.<sup>186</sup>

- 14.35 CSAM can be distributed on any platform that provides the ability to post or share images, videos or files.<sup>187</sup> This includes, for example, via instant messaging, social media, peer-to-peer networks, newsgroups, bulletin boards, and discussion forums.<sup>188</sup>
- 14.36 Hash matching is used to automate the detection of CSAM, by identifying content that matches content previously classified as CSAM and which is stored as a 'hash' in a hash database. Such content may be referred to as 'known CSAM'. Appropriate action can then be taken to remove detected CSAM, or prevent it from being uploaded.
- 14.37 We therefore consider that hash matching can, in principle, be an effective way of reducing the prevalence and dissemination of CSAM on regulated user-to-user services.
- 14.38 The Register of Risk sets out the profoundly negative impact that being sexually abused as a child has on victims and survivors. In particular, analysis by the Independent Inquiry into Child Sexual Abuse found that 88% of victims and survivors reported a negative impact on their mental health.<sup>189</sup> Child sexual abuse often also has a severe impact on physical health, including as a result of physical injury, sexually transmitted infections and pregnancy.<sup>190</sup> Further, many victims and survivors report an impact on their education, ability to work and career prospects, relationships, parenting and faith.<sup>191</sup>
- 14.39 Removing CSAM online would deliver extremely important and wide-reaching benefits:
- a) Detecting and removing CSAM can disrupt offending and lead to investigative action against those sharing and viewing CSAM online. In addition, given the connection between viewing and sharing CSAM, and committing other sexual offences against children, this is likely to surface perpetrators who are also inflicting other forms of child sexual abuse, including contact abuse.
  - b) Further, studies indicate a connection between viewing CSAM and going on to contact children for the purposes of sexual abuse. One study found that 37% of perpetrators who had viewed CSAM online went on to seek sexual contact with a child afterwards; 5% of perpetrators said that this was on a weekly basis.<sup>192</sup> This indicates that removing

---

<sup>184</sup> NCMEC is an NGO in the USA; US-based online services that find CSAM on their service are required to report this to NCMEC's CyberTipline. Where a report made to NCMEC by an online service pertains to content originating from the UK (e.g. CSAM shared by a UK user on a social media platform), NCMEC will inform the National Crime Agency (NCA). NCMEC CyberTipline Report, 2022.

<sup>185</sup> Volume 2: Chapter 6C CSEA (grooming and CSAM)

<sup>186</sup> National Centre for Missing & Exploited Children (NCMEC), 2023. [CyberTipline 2022 Report](#). [accessed 26 May 2023]

<sup>187</sup> High risk service characteristics for hosting and sharing CSAM are set out in the Register of Risk.

<sup>188</sup> Lee, H. E., Ermakova, T., Ververis, V., and Fabian, B., 2020. Detecting child sexual abuse material: [A comprehensive survey](#). [accessed 6 June 2023].

<sup>189</sup> Independent Inquiry Child Sexual Abuse, 2022. [The Report of the Independent Inquiry into Child Sexual Abuse](#). [accessed 6 June 2023].

<sup>190</sup> Independent Inquiry into Child Sexual Abuse, 2022.

<sup>191</sup> Independent Inquiry into Child Sexual Abuse, 2022.

<sup>192</sup> Insoll, T., Katariina Ovaska, A., Nurmi, J, Aaltonen, M. and Vaaranen-Valkonen, Nina., 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 12 June 2023].

CSAM on online services may also result in a reduction of other types of child sexual abuse, such as grooming and contact abuse.<sup>193</sup>

- c) Some users may inadvertently view CSAM online as a result of its wide availability. For many, this can be a traumatic experience and lead to feelings of guilt. For others, it can cause users to go on to regularly view and seek out this material, which may lead to other child sexual offences. Removing CSAM would help reduce the potential of inadvertent viewing of CSAM and these associated negative impacts.
- d) Detection of known CSAM points services and law enforcement towards communities of perpetrators or locations online where further content is being stored and shared. This leads to the discovery of unknown CSAM, which can then be hashed and added to databases to prevent its further circulation.
- e) Victims and survivors of child sexual abuse are known to experience re-traumatisation and continued re-victimisation as a result of knowing images of their abuse are circulating online, or inadvertently seeing these images. Removing CSAM on services can help to relieve this and provide reassurance to victims that online services are taking proactive measures to address the proliferation of content depicting their sexual abuse.
- f) More widespread detection and escalation of illegal content results in the identification of offenders, leading to the arrest and conviction of those possessing illegal material and/or engaging in the grooming and sexual abuse of children. It also enables the identification of victims who can then be safeguarded or protected.

## Options

14.40 We have considered whether to include in our CSEA Code of Practice an option that services deploy hash matching technology to proactively identify known CSAM.

14.41 Annex 15 considers what an effective hash matching option could look like, including:

- The type of hash matching technology (i.e., either perceptual, cryptographic, or both)
- The hash database used by services;
- The breadth of content that is scanned (and when) on the service (i.e., scanning for new content or for all existing content);
- What provision should be made about the technical performance of the technology; and,
- The use of human review in relation to content identified by the hash matching process.

## Outline measure

14.42 In light of that assessment, our proposed measure includes the following features:

- a) The use of perceptual hash matching technology to analyse content in the form of images or videos which is communicated publicly on the service. This includes both analysing content that is already present on the service (within a reasonable time), and

---

<sup>193</sup> Insoll, T., Katariina Ovaska, A. and Vaaranen-Valkonen, N., 2021. [CSAM Users in the Dark Web: Protecting Children Through Prevention \(suojellaanlapsia.fi\)](#) [accessed 14 June 2023].

content that is generated on, uploaded to or shared on the service (or that a user seeks to generate, upload or share), before or as soon as practicable after it can be encountered by other users.

- b) Comparing that content (using a suitable perceptual hash function) to an appropriate hash database of known CSAM. To be appropriate, the database should include hashes of CSAM sourced from an organisation or person with expertise in the identification of CSAM, and arrangements should be in place to ensure the accuracy of CSAM in the database (including adding hashes to the database, reviewing hashes and removing them if appropriate, and securing the database against security compromises through attacks by bad actors) and to regularly update the database.
- c) Ensuring that the technology is configured so that its performance strikes an appropriate balance between precision and recall, and reviewing this at least every six months, taking into account:
  - i) the risk of harm relating to CSAM, as identified in the service's latest illegal content risk assessment, and including information reasonably available to the provider about the prevalence of CSAM on the service;
  - ii) the proportion of content detected as a match by the technology that is a false positive; and
  - iii) the effectiveness of the systems and processes used to identify false positives.
- d) Service providers would also need to ensure a written record is made of how this balance has been struck.
- e) Ensuring that human moderators are used to review an appropriate proportion of content detected as CSAM by the technology, taking into account the principles that:
  - i) the resource dedicated to review of detected content should be proportionate to the degree of accuracy achieved by a service's perceptual hash matching technology and any associated systems; and
  - ii) this resource should be targeted at content with a higher likelihood of being a false positive.
- f) Appropriate measures should be put in place to swiftly take down (or prevent from being uploaded etc.) content detected by the technology that is correctly identified as CSAM.
- g) Service providers would need to ensure that a written record is kept of their policy for review (setting out the proportion of content they intend to review, and how the principles above have been taken into account), and keep statistical records about content reviewed.

14.43 Further detail about each of these elements is set out in Annex 15. Our provisional view is that that this would be an effective measure which has sufficient clarity for providers while also providing an appropriate degree of flexibility as to how it is adopted.

14.44 We now turn to consider:

- a) the degree of accuracy, effectiveness and risk of bias from hash matching technology deployed in accordance with the outlined option;
- b) the extent of any interference with users' rights to freedom of expression and privacy from such an option; and
- c) the costs of such an option.

- 14.45 Before considering whether it would be proportionate to include the measure in our CSEA Code and, if so, to which U2U services it would be proportionate to apply the measure.

## **Accuracy, effectiveness and lack of bias**

- 14.46 As outlined in Chapter 11, we are required to have regard to the degree of accuracy, effectiveness and lack of bias of any proactive technology we recommend.
- 14.47 There is a mature ecosystem around perceptual hash matching technology, and many larger services and some smaller services already use this technology to tackle CSAM on their services. These efforts have resulted in the identification of large volumes of CSAM, as indicated by reports to NCMEC. However, there is limited publicly available information about the uptake of hashing technology across the wider online ecosystem, especially for smaller services and certain service types.
- 14.48 Evidence suggests that hash matching technology is effective in identifying, and, by extension, facilitating the removal of, known CSAM content. Analysis of NCMEC data found that a major contributor to the exponential growth in reports made to its CyberTipline since 2009 is the rise of proactive, automated detection efforts, such as perceptual hash matching tools.<sup>194</sup> Between 2010 and 2022, the number of reports to the CyberTipline has risen from just over 10,000 per year, to over 32 million; the vast majority of these reports were generated by automated perceptual hash matching technologies.<sup>195</sup> Whilst these statistics do not translate exactly to the amount of individual pieces of CSAM detected, they do provide a strong indication that hash matching is effective in detecting CSAM.
- 14.49 As we have discussed above, the presence of CSAM online causes very significant and egregious harm. By facilitating the detection and removal of known CSAM, widespread deployment of perceptual hash matching would therefore deliver very significant benefits.
- 14.50 We understand that third-party entities support perceptual hash matching, and it forms the basis of many in-house solutions developed by larger service providers. Some services discuss their use of perceptual hash matching technology and solutions publicly, such as through transparency reporting.
- 14.51 There is limited publicly available evidence on the accuracy of perceptual hash functions. As outlined in Annex 15, the accuracy of a given perceptual hash matching technology depends on the technical parameters which have been selected, particularly the similarity threshold. The broad consensus from industry is that hash matching is the most effective and scalable means of detecting known CSAM. Statements provided to us by third-party solutions providers and services deploying in-house hash matching solutions indicate that regular review and refinement of these technical parameters, based on insights from regular system audits and feedback from human oversight, can be effectively utilised to fine-tune the accuracy of hash matching systems in their entirety. The statements provided indicate that this process of fine tuning can be used to configure hash matching systems in such a way that they are suitably accurate for the purpose intended.
- 14.52 The accuracy of the output of a hash matching process will be affected by the quality of the hashing database. The measure outlined above sets out steps that services should take to mitigate the risk of inaccuracies, as further discussed in Annex 15. We also note that its

---

<sup>194</sup> Farid, H., 2021. [An Overview of Perceptual Hashing](#). [accessed 8 May 2023].

<sup>195</sup> Fahid, 2021.

accuracy may be limited by changes made to the file to evade detection, which go beyond the detection capability of the hashing function.

- 14.53 Perceptual hash functions are subject to inherent limitations and may be vulnerable to security compromises through attacks by bad actors which have the potential to impact the accuracy of the hash matching process.<sup>196</sup> We consider that the safeguards set out in the measure outlined above and discussed at Annex 15 should mitigate the risks of inaccuracies arising from such limitations and security compromises. Further, we are aware of recent research that has indicated perceptual hashing algorithms could be repurposed to add hidden secondary capabilities.<sup>197</sup> The option we have outlined relates to the use of perceptual hashing solely for the purposes of CSAM detection.
- 14.54 We consider that traditional perceptual hash functions themselves are less likely to have inherent biases as they do not predict samples to apply to a wider population in the way that some technologies, such as machine learning, do. We consider that the main risks of bias occur with the compilation of the hash databases used, rather than the technology itself. For example, addition of hashed CSAM images to the list depends on where the original image was found online, how it is detected (e.g. through AI machine learning models, web crawling, or human analysts), and the assessment of content as CSAM (e.g., age determination). This may create biases that underrepresent the scale and nature of the problem of CSAM for different ages and minority groups. To help mitigate the risk of bias, this option provides for the hash database to include CSAM hashes sourced from a third party with expertise in the identification of CSAM.

## Costs and risks

- 14.55 As we are considering the option of applying perceptual hash matching for content communicated publicly, this subsection covers costs associated with implementing and maintaining a perceptual hash-matching system.<sup>198</sup> Based on the discussion above, we assume the measure would be deployed using an appropriate hash database and appropriate technical parameters. In addition, the measure includes the use of human moderators to help ensure that the technology operates accurately. We include these factors in our discussion of costs below. See Annex 14 for more detail on how we have quantified the costs (and some benefits) of the measure.
- 14.56 Costs involve both one-off set up costs and ongoing maintenance and operating costs. One-off costs include labour costs related to building a hash matching system. Ongoing costs include:
- a) Costs of maintaining a hash-matching system;
  - b) Cost of software, hardware and data; and,
  - c) Cost of reviewing matches, moderating content and reporting CSAM.

---

<sup>196</sup> Ofcom Overview of Perceptual Hashing paper, 2022.

<sup>197</sup> Jain, S., Cretu, A., Cully, A., and de Montjoye, Y., 2023. [Deep perceptual hashing algorithms with hidden dual purpose: when client-side scanning does facial recognition](#). [accessed 28 June 2023].

<sup>198</sup> Costs are estimated for a service that does not currently use perceptual hash matching. Where a service already has this technology in place, the costs may be less material (though this will depend on the current set-up of the tool). This option would also remove the flexibility for services who currently implement this technology to stop implementing it.

## One-off costs to build the hash matching system

14.57 The one-off cost of building a hash-matching system will largely come in the form of 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 affects these development costs. For example, we understand that it is cheaper for services to integrate hash matching via an API than by building an in-house hash-matching system. The technical complexity of the service also affects these costs, as integration will be more challenging for more complex services with a multitude of functions, for example. Larger services may have more complex operational structures and a greater number of individuals involved in making changes to a service, which can increase the resource required to implement a new technology. Through our engagement with industry experts, we understand that undertaking the initial set-up of a CSAM hash-matching system can take 2 to 18 months of full-time work by a software engineer working alongside professionals from a range of occupations. We estimate one-off set-up costs to be between £16,000 and £319,000 depending on the size and complexity of the service.<sup>199</sup>

## Ongoing maintenance costs

14.58 Ongoing costs include the labour costs of maintaining the hash-matching system. Activities include applying updates, adjusting parameters, and ingesting new hash lists, and integrating with new products. 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.<sup>200</sup> As with the cost of building a hash-matching system, the cost of maintaining the system is likely to scale with the size of the service, as larger services are generally more complex and therefore 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 £80,000 depending on the size and complexity of a service.

## Ongoing software, hardware and data costs

14.59 Ongoing costs will also include the annual cost of software, hardware and data. These costs will generally be larger for services with a large user base because it is common practice for NGOs in the hash-matching industry to charge based on the capacity of the service to pay for their product. Larger services, especially those with more complex product portfolios, will also more often opt for in-house solutions that require multiple hash lists and software products.<sup>201</sup> Although lower-cost solutions are available, we assume that the annual cost of software, hardware and data starts at £25,000<sup>202</sup> and can go into six-figure sums for the largest services.<sup>203</sup>

---

<sup>199</sup> This range has been constructed by modelling services that reach 70,000 to 7 million UK users. See Annex 14 for more detail. 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, not least because some services will be larger than those that have been modelled to obtain the upper estimate.

<sup>200</sup> Common assumptions on costs are detailed in Annex 14.

<sup>201</sup> For example, IWF and Thorn currently charge services based on capacity to pay, number of API queries, and other considerations.

<sup>202</sup> In practice, these costs may be lower where a service is deemed by a relevant NGO provider to have less ability to pay.

<sup>203</sup> These assumptions are based on our own expertise and industry experts. See Annex 14 for more detail.



## Ongoing content moderation costs

14.60 The largest ongoing cost is likely to be labour costs related to content moderation.<sup>204</sup> In general, these costs will scale with the amount of content on a service as, all else being equal, these services will experience more positive hash matches that need to be reported to the relevant authorities and potentially subject to human review (with the potential for more user complaints). The costs will also be higher for services with more known CSAM. As detailed in Annex 14, we have therefore assumed that moderation costs depend on the size and kind of service, where the kind of service relates to the service's risk of CSAM being present on the service. The analysis, which is based on our own expertise and engagement with industry experts, finds that services with a small user base and a relatively low risk of CSAM won't require a full-time moderator, whereas high reach/risk services will likely employ a small team of moderators. For example, we estimate that a high-risk service with 700,000 active UK users per month could spend between £21,000 to £61,000 on moderators per year. As moderation costs are broadly proportionate to the amount of CSAM on a service, we would expect services to experience large moderation costs only if there are large benefits to adopting the measure.

---

<sup>204</sup> 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.

14.61 We are aware that, in addition to the above costs, there may be other costs to the service provider as well as costs (both monetary and non-monetary) to other organisations and individuals.<sup>205</sup>

## Total costs

14.62 Based on the costs discussed above, and following the assumptions outlined in Annex 14, we estimate that one-off costs could range from £16,000 for a small service (a service that reaches 70,000 UK users per month) to £319,000 for a large service (a service that reaches 7 million UK users per month). We estimate that the ongoing annual costs could range from £31,000 annually for a service that reaches 70,000 users to £820,000 for a service that reaches 7 million users which has a high volume of CSAM on the service and therefore requires a team of moderators to review and report the content.

14.63 Table 14.1 summarises these findings. Further detail on total costs for different sizes and kinds of service are included in Annex 14.

**Table 14.1: Illustrative range of costs for services implementing the hash-matching measure**

Lower estimate: Low reach (70,000 UK users)		Upper estimate: high reach (7 million UK users)	
Build cost (one-off)	Ongoing costs (annual)	Build cost (one-off)	Ongoing costs (annual)
£16,000	£31,000	£319,000	£820,000 <sup>206</sup>

Source: Ofcom analysis (see Annex 14 for more detail)

14.64 Our analysis suggests that the cost of implementing and maintaining a hash-matching system is greatest for the largest and most complex services, and that costs would be expected to be significantly lower for smaller and simpler services.

14.65 Even though the costs of the measure are likely to be smaller for services with a small user base, we are mindful that the cost of hash matching may represent a larger proportion of their revenue. Low-capacity services such as these may also experience additional barriers to undertaking hash matching, such as a lack of specialist policy knowledge about CSAM or pre-existing engineering expertise to integrate perceptual hash matching across their products.

14.66 Nevertheless, we understand that some smaller services are already using perceptual hash matching to identify known CSAM, which suggests that these barriers are surmountable for services with small user bases. Indeed, our research found that the kind and size of services that deploy hash matching is varied, and includes many different kinds as well as different sizes of service.

<sup>205</sup> Services would also need to put in place protocols for those dealing with the hashing system to ensure that the material is handled safely, and we recognise that this will incur some costs for services. Actions taken could include, for example, ensuring that only named individuals are permitted access and involved with the implementation/testing/review of the system.

<sup>206</sup> This is modelled on a high-risk service and so the on-going costs are higher than for the high-reach medium-risk service presented in Table A14.4 in Annex 14.

## Dynamic effects on the market

- 14.67 Hash matching is a relatively mature technology that is already being used by many of the biggest internet services (and some smaller services) to detect known CSAM. However, there could be challenges with the capacity of current database providers to support a significantly more widespread use of hash matching in a short time frame.
- 14.68 We recognise that there may however be benefits for the market should there be a wider deployment of hash matching technology. In particular, an increase in the number of internet services that implement hash matching to detect known CSAM would likely have a positive impact on the level of interest, investment, and innovation in the use of perceptual hash matching to reduce known CSAM online. A welcome impact of this would be a likely reduction in the cost of adopting hash matching and an improvement in the effectiveness of hash matching systems.
- 14.69 Further, it is also important to note potential risks involved in the more widespread adoption of hash matching technology. One response by perpetrators might be to manipulate content in an attempt to evade matching algorithms. This is why our provisional view is that perceptual hashing is likely to be a more effective hash matching technology; it can reduce this risk, as the balance between precision and recall can be adjusted in response to trends in the way that content is distorted. Another response might be for perpetrators to avoid internet services that deploy hash matching technology. We consider this as part of our discussion in paragraph 14.104 below about the services that any CSAM hashing recommendation should apply to.

## Rights impacts

- 14.70 As set out in Chapter 12, content moderation functions can have particular impacts on individuals' and entities' rights to privacy under Article 8 of the European Convention on Human Rights ('ECHR') and to freedom of expression under Article 10 ECHR. Our focus in this sub-section is on the impacts to users' rights relating to posting of content that is not CSAM, and in particular potential interferences arising from users' content being wrongly identified as CSAM.<sup>207</sup>
- 14.71 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, and an interference with the right to freedom of expression must be prescribed by law and necessary in a democratic society in pursuit of a legitimate interest. In either case, in order to be 'necessary', the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.
- 14.72 In considering whether impacts on these rights are proportionate for the purposes of this measure, our starting point as set out in Chapter 12 is to recognise that Parliament has determined that CSAM should constitute priority illegal content under the Act, and that a substantial public interest exists in measures which aim to reduce its prevalence and dissemination online. That public interest relates to each of the prevention of crime, the protection of health and morals, and the protection of the rights of others.
- 14.73 The detection, removal and reporting of CSAM acts directly to prevent crime in a number of ways, such as by deterring users from posting such illegal content or by preventing other

---

<sup>207</sup> We consider that any interference with users' rights that related to content correctly identified as CSAM would clearly be proportionate.

users from accessing it (and so potentially committing further offences). It similarly acts to protect public morals, including by preventing users inadvertently encountering CSAM online.

- 14.74 The removal of CSAM also acts directly to protect the rights of victims of child sexual abuse. CSAM causes ongoing harm to victims from knowing that material depicting their sexual abuse 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' rights under Article 8 ECHR and protects victims' personal data. We place considerable weight on these positive impacts (consistent with our statutory duty to have regard to the vulnerability of children under section 3(4)(g) of the Communications Act 2003).
- 14.75 Further, we explained above that there is evidence that measures that restrict the dissemination of CSAM online may reduce levels of child sexual abuse, both by reducing contact offending associated with viewing CSAM and by enabling offenders who may be involved in contact offending to be identified (through reporting of CSAM to law enforcement authorities). As well as the prevention of crime, any such reduction in child sexual abuse would protect the fundamental values and essential aspects of private life in relation to children (including their health), as well as children's rights not to be subjected to inhuman or degrading treatment under Article 3 ECHR. 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.
- 14.76 We turn now to consider adverse impacts on users' privacy and freedom of expression.

## Users' privacy

- 14.77 The hash matching option under consideration would involve the proactive monitoring of relevant content to detect CSAM. However, consistent with the Act's constraints on proactive technology, this option only specifies that hash matching occurs of content that is communicated publicly by means of the service. As such, it should not affect the confidentiality of communications. However, as explained in our draft guidance on when content should be considered to be communicated publicly or privately, in Annex 9, we consider that there could be some circumstances in which a person could still have a reasonable expectation of privacy (for the purposes of Article 8 ECHR) in relation to content which is nonetheless considered to be communicated publicly for the purposes of the Act. In addition, the processing of images and videos may involve the processing of personal data of individuals.
- 14.78 Insofar as that processing is limited to the automated use of an algorithm to create a hash from the content for the purpose of hash matching, any interference with users' rights to privacy under Article 8 ECHR would be slight. Such processing will also need to be undertaken in compliance with relevant data protection legislation (including, so far as the UK GDPR applies, rules about processing by third parties or international data transfers).
- 14.79 Review of detected content by human moderators, including those employed by a contracted third party, involves more significant potential impacts on privacy. The impact on users generally would be limited, as the moderators would only access content that the user has already communicated publicly. There is a more substantial impact on the privacy of victims depicted in detected CSAM which is reviewed by moderators. However, that review (and the associated interference) is for the purpose of ensuring content is taken down

accurately, and thus important to limiting impacts on users' freedom of expression. As explained below, we consider human review to be a necessary part of ensuring the overall measure – which acts to *protect* victims, including from serious potential impacts on their privacy of the further dissemination of CSAM – is proportionate (and do not consider a less intrusive approach (as regards victims' privacy) would be sufficient).

- 14.80 Interference with users' or other individuals' privacy rights may also arise insofar as the option would lead to reporting to reporting bodies or other organisations in relation to CSAM detected using perceptual hash matching technology – for example, that a user was responsible for uploading content detected as CSAM to the service.
- 14.81 In particular, section 66 of the Act makes provision which (when brought into force) will require providers of regulated user-to-user services to report detected and unreported CSEA content to the Designated Reporting Body housed in the NCA (as further specified in the Act and to be specified in regulations made by the Secretary of State under section 67 of the Act). Providers may also have obligations to report CSEA content in other jurisdictions, or may have voluntary arrangements in place. For example, US providers are obliged to report to NCMEC under US law when they become aware of child sexual abuse on their services.
- 14.82 In part, any such interference results from the duties created by the Act or by existing legislation in other jurisdictions. In particular, where users or other individuals are correctly reported pursuant to the Act because they are suspected of committing the offence related to the CSEA content, any interference with their rights 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.
- 14.83 However, we have considered the extent to which the inclusion of this measure in our Code of Practice as a recommended measure for the purpose of complying with providers' illegal content safety duties might give rise to additional interference.
- 14.84 As explained above, use of perceptual hash matching can result in cases where detected content is a false positive match for CSAM, or a match for content that is not CSAM and has been wrongly included in the hash database. These cases could result in individuals being incorrectly reported to reporting bodies or other organisations, which would represent a potentially significant intrusion into their privacy.
- 14.85 It is not possible to assess in detail the potential impact of incorrect reporting of users: the number of users potentially affected will depend on how services implement hash matching; while further details of the reporting requirements under the Act are to be specified by the Secretary of State in secondary legislation. However, the option includes principles and safeguards in relation to the hash database, the configuration of the technology, and the use of human moderators that are designed to help secure that the technology operates accurately. We discuss these further below.
- 14.86 In addition, reporting bodies have processes in place to triage and assess all reports received, ensuring that no action is taken in cases relating to obvious false positives. These processes are currently in place at NCMEC and will also be in place at the Designated Reporting Body in the NCA, to ensure that investigatory action is only taken in appropriate circumstances.

## Users' freedom of expression

- 14.87 Potential interference with users' freedom of expression arises insofar as content is taken down on the basis of a false positive match for CSAM or of a match for content that is not CSAM and has been wrongly included in the hash database. In addition, there could be a risk of a more general 'chilling effect' if users were to avoid use of services which have implemented hash matching in accordance with our option. However, we do not consider that any such effect would be significant, given that 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 it on to relevant national authorities including the NCA.
- 14.88 As mentioned above, the design of this option includes important safeguards that mitigate against the risk that content is wrongly identified as CSAM. In particular, it includes principles in relation to use of an appropriate hash database, including the need for governance arrangements to be in place that ensure (so far as practicable) that content is added to the hash database correctly, and allows for hashes to be reviewed and removed if found to be wrongly included, and for the database to be secured from security compromises through attacks by bad actors. It also includes principles relating to the configuration of the technology to ensure that the hash matching technology strikes an appropriate balance between precision and recall, taking into account (among other things) the proportion of content detected as a match by the technology which is a false positive, and the effectiveness of the systems and processes used by the service to identify false positives. These principles further provide for the performance of the system, and whether the balance between precision and recall continues to be appropriate, to be reviewed at least every six months. These principles aim to help ensure the accuracy and effectiveness of the technology, and should therefore limit the risk of detection and reporting of non-CSAM content.
- 14.89 However, the measure's design also provides flexibility in several respects, including to choose an appropriate hash database and as to the configuration of the perceptual hash matching. As a result of this, there could be variation in the impact on users' freedom of expression arising from services' different implementations of perceptual hash matching, and implementations that substantially impact on freedom of expression could be in accordance with the measure in our Code of practice.
- 14.90 Services' different implementations of the measure will be influenced by their commercial and reputational incentives, but we would not expect these to point in one direction. Services have incentives to minimise the amount of CSAM on their platforms, but also to limit the amount of content that is wrongly taken down (to avoid creating friction for users), or to reduce the costs of human moderation or operating complaints procedures.
- 14.91 Impacts on freedom of expression could in principle arise in relation to the most highly protected forms of content, such as religious or political expression, and in relation to kinds of content that the Act seeks to protect, such as content of democratic importance and journalistic content. However, we consider there is unlikely to be a systematic effect on these kinds of content: for instance, such content would be unlikely to be particularly vulnerable to false positive detection. In particular, we consider that this option would be unlikely to result in the disclosure of journalistic sources to law enforcement authorities, given that the measure concerns exclusively images and videos and only applies to content communicated publicly.

- 14.92 To further mitigate potential impacts on freedom of expression (and associated impacts on privacy), this option provides that services should review an appropriate proportion of content detected as a match for CSAM, determined taking into account specified principles, and should keep a written record of their policy for review of detected content. Human moderation of detected content can act as an important safeguard to confirm that content detected by hash matching is indeed CSAM, before it is taken down.
- 14.93 We recognise, however, that this only acts to mitigate the impact on freedom of expression in relation to the proportion of content that is reviewed (which is not specified in the safeguard). We further recognise that services are not required to adopt the safeguard and are permitted to take alternative approaches to complying with their duty about freedom of expression under section 22(2) of the Act (which requires them not to achieve particular outcomes but only to have particular regard to the importance of protecting users' right to freedom of expression within the law when deciding on and implementing safety measures and policies).
- 14.94 It should be noted that this option uses hash matching to detect CSAM, but recognises that it is open to service providers to use hash matching to detect content other than CSAM (subject to complying with their other duties under the Act, including about freedom of expression). This option makes clear that services taking this approach should ensure that the principles and safeguards included in the measure to promote accuracy in relation to CSAM remain effective: for instance, by reviewing an appropriate proportion of all detected content if they are unable to distinguish between content detected as CSAM and as other hashed material. (So far as content other than CSAM may be wrongly taken down as violative of a service's terms of service, this falls outside the scope of this analysis.)
- 14.95 Where a service takes down content on the basis that it is illegal content, complaints procedures operated pursuant to section 21(2) of the Act allowing for the relevant user to complain and for appropriate action to be taken in response may also mitigate the impact on users' rights to freedom of expression.<sup>208</sup>

## Provisional conclusion

- 14.96 CSAM is recognised in the Act as priority illegal content, and regulated user-to-user services are therefore under a statutory duty to take proportionate measures (including in relation to content moderation) to prevent individuals from encountering that content by means of the service and use proportionate systems and processes to minimise the length of time for which it is present. Our own assessment set out in our Register of Risks has also reinforced the severity of the harm caused by the dissemination of CSAM, and indicates the very high prevalence of CSAM online.
- 14.97 Our provisional view is that perceptual hash-matching technology can be an effective content moderation tool - when deployed using a high-quality hash database and appropriate technical parameters – to help services proactively identify known CSAM on their service at pace and at scale. It is an established technology that is widely used in the online safety sphere and facilitates the removal of millions of items of CSAM each year.
- 14.98 We recognise, however, that its use can impose costs on services and the use of proactive technology can potentially interfere with users' rights to freedom of expression within the law and privacy. However, the severity of the harm we would be addressing, and the

---

<sup>208</sup> See Chapter 16 Reporting and complaints.



benefits that would result from reducing this, means that we provisionally consider it would be proportionate to recommend that at least some services deploy hash matching for known CSAM.

- 14.99 We have carefully considered to which services it would be appropriate to apply the measure outlined above in the CSEA Code of Practice. That measure is discussed in more detail at paragraphs A15.2-A15.56 of Annex 15, and set out in draft in Annex 7.
- 14.100 To inform our view, and as explained in Chapter 11, we have had regard to the principle that measures should be proportionate, taking into account both:
- a) our assessment of the risk of harm presented by services of different kinds and sizes (including as to levels of risk and as to the nature, and severity, of potential harms to individuals); and
  - b) the size and capacity of providers of services.
- 14.101 **Services have different levels of risk for CSAM.** Our draft Service Risk Assessment Guidance, set out in Annex 5, recognises that not all services will have the same risk of CSAM and sets out the factors which we consider indicate that a service should be assessed as being at low, medium or high risk for CSAM, including image-based CSAM. These are based on the risk factors set out in the Register of Risk and Risk Profiles, and on the available evidence, which suggests that those factors can facilitate the sharing of CSAM or may be sought out by perpetrators seeking to share CSAM.
- 14.102 The guidance sets out our provisional view that a service is likely to be at **high risk** for image-based CSAM if it allows images or videos to be uploaded, posted or sent and:
- a) The service has systematically<sup>209</sup> been used by offenders to upload image-based CSAM; or
  - b) any of the following apply:
    - i) The service has a majority of<sup>210</sup> relevant risk factors associated with CSAM in Ofcom’s Risk Profile for U2U services, in addition to allowing images or videos to be uploaded, posted or sent;
    - ii) The service is a file-storage and file-sharing service;<sup>211</sup>
    - iii) The service is an online adult service, or a service which allows pornographic content;

---

<sup>209</sup> This will be based on previous pattern of use by offenders to upload and share content. For further information, see Annex 5, Service Risk Assessment Guidance.

<sup>210</sup> Risk factors for image-based CSAM includes: child users; social media services; private messaging services; discussion forums and chat rooms; user groups or group messaging; livestreaming; direct messaging; encrypted messaging. For further information, see Annex 5, Service Risk Assessment Guidance.

<sup>211</sup> For the purposes of this measure a file-storage and file-sharing service is to be 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 proposed definition is consistent with that used in the risk profiles and the Register of Risk. The utilisation of link-based sharing of content which is stored online as an integral component of its specialised functionality is what sets these platforms apart from other more generalised online services which allow general sharing of content. This encompasses services offering specialised image hosting and link sharing, and which may also allow for the images to be embedded across other services. For further information, see Annex 5, Service Risk Assessment Guidance.

- iv) The service allows users to upload, post or send content without creating an account.<sup>212</sup>

14.103 Our provisional view is that a service is likely to be at **medium risk** for image-based CSAM if it does not meet the criteria for being at high risk and it allows images or videos to be uploaded, posted or sent **and**:

- a) There is evidence that the service has been used by offenders to upload, post or send image-based CSAM recently; **or**
- b) The service has two or more<sup>213</sup> relevant risk factors<sup>214</sup> in the Risk Profiles associated with CSAM, in addition to allowing images or videos to be uploaded, posted or sent.

14.104 Services' illegal content risk assessments offer a potential means to target the measure at services with higher risk (where the measure carries higher potential benefits). However, we recognise that this depends on services' conducting robust risk assessments, in accordance with their statutory duties.

14.105 **We propose to target our measure based on a service's risk according to their user reach.**

When considering which services a measure on hash matching for image-based CSAM should apply to, we considered four broad options:

1. Recommend that all services deploy perceptual hash matching to detect known CSAM;
2. Recommend that all services whose risk assessment identified them as at least a low risk for image-based CSAM deploy perceptual hash matching;
3. Recommend that all services whose risk assessment identified them as a medium or high risk for image-based CSAM deploy perceptual hash matching to detect known CSAM;
4. Recommend adoption of perceptual hash matching for image-based CSAM for the following services:
  - i) Services with medium-risk for image-based CSAM in their most recent illegal content risk assessment and more than 7 million monthly UK users;
  - ii) Services with high risk for image-based CSAM in their most recent illegal content risk assessment and more than monthly 700,000 UK users; and
  - iii) Services that are high risk for image-based CSAM and which are file-storage and file-sharing services with more than 70,000 monthly UK users.

---

<sup>212</sup> This refers to services which allow users to post or send content without the need to register (for example with an email address) or to provide any log-in details; this can result in a user's identity being unknown (or partially unknown) to a service, as well as other users. This does not refer to 'pseudonymous' users, where a person registers for a service but does not necessarily use any personally-identifying information. This is also distinct from simply using a service without logging in (such as browsing a webpage), and specifically refers to the ability to post or send content on the service. For further information, see Annex 5, Service Risk Assessment Guidance.

<sup>213</sup> Risk factors other than allowing images or videos to be uploaded, posted or sent. For further information, see Annex 5, Service Risk Assessment Guidance.

<sup>214</sup> Child users; social media services; private messaging services; discussion forums and chat rooms; user groups or group messaging; livestreaming; direct messaging; encrypted messaging. This list excludes risk factors from the Risk Profiles which are indicated in the 'high risk' criteria. For further information, see Annex 5, Service Risk Assessment Guidance.

- 14.106 We rejected the first option on the basis that some U2U services do not include functionalities to allow users to post or send images or videos. Such services would have no risk related to users sharing image-based CSAM, and it would not be relevant to recommend they implement hash matching of user-generated content to detect image-based CSAM.<sup>215</sup> Similarly, we rejected the second option on the basis of our provisional view that it would not be proportionate to recommend this measure to services who assess as low-risk<sup>216</sup> in their risk assessment for image-based CSAM. This is for the following reasons:
- 14.107 As we set out above, services implementing hash matching may incur significant costs. We are not satisfied that it would be proportionate to recommend services which have a low risk of image-based CSAM incur those costs.
- 14.108 We are concerned that extending this measure to low-risk services could put significant pressure on the wider hash-matching ecosystem and the ability of relevant providers to meet a significant increase in demand in the short-term. This could result in barriers to services being able to act in accordance with the measure.
- 14.109 Finally, although not extending this measure to low-risk services may risk displacement of CSAM posting and sending to these services, Ofcom recommends that risk assessments are reviewed at least every 12 months. If displacement were to occur to low-risk services, on reviewing their risk assessment, these services should assess themselves as now being at medium or high risk for image-based CSAM and the measure would then apply to them.
- 14.110 We also rejected the third option to apply the measure to all services identified as medium or high-risk for image-based CSAM. Given the costs associated with hash matching and the difficulty some services may have in bearing these costs, it is not clear that it would be proportionate to propose this measure for all medium-risk services regardless of the size of their user base, as the benefits of the measure are likely to be greater for medium-risk services with a higher reach. In addition, there are only a finite number of providers of hash matching databases and we understand these services only have a finite amount of capacity to provide their services to clients. Therefore, we are concerned that if we recommended the measure for all medium and high-risk services, regardless of size, the database ecosystem would not be able to cope with levels of demand in the short term.
- 14.111 We consider that the fourth option would be proportionate. Given the financial resources typically available to them and the scale of the benefits associated with the measure, we consider that it is proportionate for large medium-risk services to deploy hash matching for known CSAM. Given the severity of the harm and the benefits in reducing it, we also consider that it is proportionate for high-risk services with 700,000 monthly users to deploy hash matching. By applying the measure only to high-risk services above this user reach, we consider that we will avoid a scenario in which so many services wish to access hash-matching databases that the database ecosystem is unable to cope with the demand in the short term and services face barriers to adopting this measure.

---

<sup>215</sup> This does not mean that these services do not have risk for the sharing of CSAM URLs; in this case services with no or negligible risk for image-based CSAM may be at medium or high risk for sharing CSAM URLs and may fall into scope of the CSAM URL detection measure.

<sup>216</sup> Low risk services would include services which allow users to post images and videos, but do not have any other risk factors associated with image-based CSAM other than image/video sharing in their risk profile (or which have already adopted measures that demonstrably ensure CSAM is highly unlikely to occur on the service).

- 14.112 In addition, we are proposing to set a lower user reach threshold for file-storage and file-sharing services and bring these services in scope of the measure if they have a reach of more than 70,000 monthly UK users. As is outlined in the Register of Risk, there is strong evidence to suggest that file-storage and file-sharing services (including image sharing services) are particularly high-risk for image-based CSAM.<sup>217</sup> In 2021, INHOPE found that over half of CSAM reported to INHOPE hotlines was found on file or image hosting sites (26% and 25%, respectively).<sup>218</sup> File-storage and file-sharing services therefore play a more central role in the dissemination of known CSAM than other types of services we are aware of. We consider that this warrants us taking a more targeted approach to this type of service.
- 14.113 We also note that at the time of writing, only a handful of file-storage and file-sharing services reach more than 700,000 monthly UK users. If we set the same user number threshold for file-storage and file-sharing services as for other types of high-risk service, this would mean that the proposed hash-matching measure would not apply to many file-storage and file-sharing services known to be high-risk. Given the available evidence about the role that file-storage and file-sharing services play in hosting CSAM, there is a danger that this could leave a material part of the CSAM threat unaddressed.
- 14.114 Whilst we recognise that the costs of hash matching may represent a larger share of revenue for smaller service providers than for larger service providers, we do not consider this by itself a reason to not apply the measure to the smaller services we propose to bring in scope. As we explained above, hash matching can lead to safeguarding of child victims. While it is challenging to put a monetary estimate on the severe harm caused by CSAM, a study by the Home Office in 2019 estimated that the value of safeguarding one child from CSA was £89,240, which would be £101,700 in 2022 prices, accounting for inflation.<sup>219</sup> This demonstrates how significant the harm is, and that even if the number of child safeguards resulting from hash matching are small, the benefits can be very large. As outlined above, the benefits of hash matching go beyond safeguarding and include benefits such as reductions in re-traumatisation and re-victimisation and fewer people inadvertently viewing CSAM. Moreover, our evidence suggests that costs are likely to be lower for smaller and simpler services, and that some smaller services are already implementing perceptual hash matching.
- 14.115 As illustrated by the analysis presented in Annex 14, hash matching could be proportionate for many services accounting for the potential benefit from safeguarding alone, including the low reach services we propose to bring in scope of the measure. While there are limitations to this type of analysis, this supports our provisional conclusion that even when a limited range of benefits are considered, these are so significant that they are greater than the costs even for a number of small services.

---

<sup>217</sup> See Volume 2: Chapter 6C CSEA (grooming and CSAM).

<sup>218</sup> INHOPE Association, 2021. [Annual Report 2021](#). [accessed 15 June 2023].

<sup>219</sup> This estimates the financial and non-financial (monetised) costs relating to all children who began to experience contact sexual abuse, or who continued to experience contact sexual abuse, in England and Wales in the year ending 31st March 2019. Accounting for inflation, the total cost per victim of £89,240 would be £101,700 in 2022 prices. Source: Home Office (Radakin, F., Scholes, A., Soloman, K., Thomas-Lacroix, C., Davies, A.), 2021. [The economic and social cost of contact child sexual abuse](#). [accessed 14 August 2023].

## Summary

14.116 Taking account of the risk to users and other individuals, and the severity and nature of the harm associated with its dissemination, our provisional view is that it is proportionate to recommend the use of perceptual hash matching for CSAM in our CSEA Code of Practice in relation to the following services:

- services that reach more than 7 million monthly UK users and that are at medium or high risk of image-based CSAM;
- services that are at high risk of image-based CSAM and reach more than 700,000 monthly UK users; and
- services that are at high risk of image-based CSAM and which are file-storage and file-sharing services that reach more than 70,000 monthly UK users.

14.117 Our proposed measure has been designed with the potential impacts on users' freedom of expression or privacy in mind and incorporates a number of important safeguards to mitigate them, as outlined above. We acknowledge that the flexibility provided for in the measure could mean that hash matching adopted in ways consistent with our measure could result in some content being wrongly taken down as CSAM (noting that even technology operating with very low false positive rates could result in large numbers of false positives when applied in relation to millions of items of content).

14.118 We nevertheless consider that such impacts, including interference with users' rights to privacy under Article 8 ECHR and to freedom of expression under Article 10 ECHR, are justified by the substantial public interest in the prevention of crime, the protection of health and morals, and the protection of the rights of victims and children that this proposed measure is designed to achieve, and are proportionate to the anticipated benefits of the measure from reducing the prevalence and dissemination of CSAM. We also do not consider that there is a less intrusive way of achieving these aims.

14.119 We welcome input and evidence from stakeholders on our proposed approach, particularly on the impact on smaller services of applying this measure.

## Hash matching for terrorism content

---

14.120 We have considered whether to also propose a hash matching recommendation for the detection of image-based terrorism content.

14.121 In principle, we think that such a measure could be effective, and we recognise that a number of services are already deploying hash matching for this kind of content. However, for the reasons set out below, we are not proposing to recommend this in our first Code of Practice. We are however keen to gather evidence on this from stakeholders.

## Harms that hash matching seeks to address

14.122 In this context 'terrorism content' refers to images or videos which, when posted, uploaded or forwarded/reposted to an online service, would amount to an offence listed in Schedule 5 to the Act.<sup>220</sup>

---

<sup>220</sup> For more detail, see Chapter 2 (Terrorism) of the draft Illegal Content Judgement Guidance.

- 14.123 Terrorism content exists online across an ecosystem of platforms. Terrorist actors, like all users of the internet, use a breadth of internet sites, platforms and spaces for different purposes. For decades, terrorist actors have set out to adopt and exploit new technologies to facilitate their operations.<sup>221</sup> Increasingly we are witnessing them exploit new functionalities offered by online services, like livestreaming.
- 14.124 Recently, there has been a migration of terrorism content from ‘conventional’ social media services towards more fringe services, as a result of concerted law enforcement action from US, UK and EU governments, pressure from stakeholders to act against terrorism content, and more sophisticated moderation efforts from the largest social media platforms, like the sharing of hash databases.<sup>222</sup> This migration to smaller services is well documented.<sup>223</sup>
- 14.125 Nonetheless, terrorism content in the form of images, videos and PDFs continues to be present across different services. We understand this includes publications and propaganda videos that are instructional, ideological and inspirational.
- 14.126 The scale and pace of dissemination of such illegal content, across the ecosystem, presents a significant problem for service providers’ moderation practices.
- 14.127 The livestream of the Christchurch attack, in which 51 individuals were killed in an anti-Muslim terrorist attack in 2019, garnered approximately 4,000 views before removal. It is reported that none of the viewers of the livestream reported the video to Facebook, which received its first user report 29 minutes after the video started, and 12 minutes after the live broadcast ended.<sup>224</sup> 8chan was then allegedly used to share a link to the footage. In the 24 hours that followed, Facebook removed 1.5 million videos of the attack and a further 1.2 million were blocked at upload.<sup>225</sup>
- 14.128 The most severe harm amounting from this content is potential radicalisation, and inspiration of future terrorist attacks.
- 14.129 In a case study on UK Islamist terrorism cases, the Institute for Strategic Dialogue found that the “*overwhelming majority of Islamist terrorist content found in UK terrorism investigations*” are historical and theological texts produced by al-Qaeda.<sup>226</sup>
- 14.130 In a recent series of livestreamed terrorist attacks, imagery and motifs from prior attacks have been seen to influence those who have yet to carry out violence. Multiple expert researchers have pointed to this link, and indeed demonstrated similarities between manifestoes of attackers that precede one another, published and available online. In the aftermath of the Buffalo attack in 2022, the U.S. Department of Homeland Security, Federal Bureau of Investigation and the National Counterterrorism Center assessed that the

---

<sup>221</sup> Alrhmoura, A., Winter, C., and Kertésza, J., 2023. [Automating Terror: The Role and Impact of Telegram Bots in the Islamic State’s Online Ecosystem](#). [accessed 14 May 2023].

<sup>222</sup> 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.

<sup>223</sup> Amarasingam, A., Maher, S., and Winter, C., 2021. [How Telegram Disruption Impacts Jihadist Platform Migration](#). [accessed 13 May 2023].

<sup>224</sup> Mclure, T., 2022. [Buffalo shooting: unease in New Zealand as live stream of ‘Christchurch-inspired’ attack finds foothold](#). *Guardian*. 18 May 2022. [accessed 14 May 2023].

<sup>225</sup> Macklin, G., 2019. [The Christchurch Attacks: Livestream Terror in the Viral Video Age](#), *CTC Sentinel*, 12 (6). [accessed 15 May 2023]; Meta, 2019. [Update on New Zealand](#). [accessed 15 May 2023].

<sup>226</sup> ISD (Davey, J., Comerford, M., Guhl, J., Baldet, W. and Colliver, C.), 2021. [A Taxonomy for the Classification of Post-Organisational Violent Extremism & Terrorism Content](#). [accessed 14 May 2023].



associated manifesto will “likely enhance the capabilities of potential mass casualty shooters who may be inspired by this attack”.<sup>227</sup>

14.131 Further, there is also serious harm associated with being depicted in, or the subject of, the content if the user is a survivor of a terror attack. The Christchurch Call Unit explained that videos of both the Christchurch and Buffalo attacks were sent directly to victims of the Christchurch attacks with hateful messages.<sup>228</sup>

14.132 We consider that content on social media, video-sharing and file-storage and file-sharing services presents the most acute risk to users. In the most recent report on the Internet and radicalisation pathways, increased use of forums/chatrooms and open social media platforms was evident for all groups studied.<sup>229</sup> The latest transparency report from Tech against Terrorism’s TCAP database found that file sharing services were the most exploited technology type.<sup>230</sup>

14.133 In sum, the scale and pace of dissemination of terrorism content in the form of images, videos and PDFs present a high risk of harm when encountered by users, which can result in serious real-world consequences and a risk to life.

## Options

14.134 We have considered recommending measures to disrupt the pace and scale of dissemination of terrorism content. Some of the relevant measures that we identified included:

- The creation of a hash matching database by an in-scope service, operating internally across uploaded content communicated publicly by means of the service.
- Subscribing to a third-party hash matching database that can be deployed across an in-scope service, against content communicated publicly by means of the service. This could include the Global Internet Forum for Counter Terrorism’s (GIFCT) Hash-Sharing Database.<sup>231</sup>

14.135 Ofcom understands that there is broad take up of these measures amongst many online services (including but not limited to services that belong to GIFCT’s database), and research demonstrates the effectiveness of such tooling in moderating terrorism content at scale, and especially at pace in the aftermath of specific incidents.

14.136 However, at present, we need further evidence on a number of areas before we are in a position to propose a recommendation in respect of hash matching for terrorist content in our Code of Practice. These areas include:

---

<sup>227</sup> Sganga, N., 2022. FBI, [DHS issue bulletin warning of potential for racially motivated copycat attacks](#), *CBS News*, August 24 2022. [accessed 15 May 2023].

<sup>228</sup> O’Callaghan, J., 2022. [March 15 survivors retraumatised by link to Buffalo attack livestream video](#), *Stuff*, May 17. [accessed 15 May 2023].

<sup>229</sup> Kenyon, J., Binder, J., and Baker-Beall, C., 2022. [The Internet and radicalisation pathways: technological advances, relevance of mental health and the role of attackers](#). [accessed 15 May 2023].

<sup>230</sup> The TCAP database is made up of verified terrorist content collected in real time from messaging platforms and apps, which notifies technology companies of the presence of such content on their platforms. Most content was located on file-sharing sites (6,526 alerts sent between December 2021 and November 2022). Source: Tech against Terrorism, 2023. [Transparency Report: Terrorist Content Analytics Platform, Year Two: 1 December 2021 - 30 November 2022](#). [accessed 10 October 2023].

<sup>231</sup> GIFCT, 2023. [GIFCT’s Hash-Sharing Database](#). [accessed 15 May 2023].



- the accuracy and effectiveness of hashing solutions for terrorism content specifically (including their false positive/false negative rates);
- the extent to which a hashing solution can identify terrorism content accurately in different contexts from which the hash was derived from, with potential implications for users' freedom of expression;
- the degree of human oversight necessary to ensure the technology is sufficiently accurate in identifying terrorism content; and
- the potential costs associated with hash matching for terrorism content, including setup and maintenance (e.g. when creating an internal database, when connecting with an external provider, and in terms of moderation costs).

14.137 A key distinction here between our position on hash matching for CSAM is that we understand the terrorism offences to be quite different in nature to those relating to CSAM. In particular, we recognise that it may be more challenging for services to identify content as amounting to a terrorism offence (under the reasonable grounds to infer standard explained further in our Illegal Content Judgements Guidance) because of the importance of contextual information.

## Provisional conclusions

14.138 The dissemination of terrorism content across public channels results in harm for users and continues to be a problem for services. The deployment of hash matching could potentially address this by facilitating quick identification, review and removal of illegal content. This would help limit the dissemination of the content within a platform at pace and scale; this in turn can prevent high volumes of views and therefore reduces the risks to users.

14.139 However, we need to explore further the accuracy of such a measure and its implications on costs and potential rights impacts. Therefore we have decided not to propose the inclusion of measures relating to hash matching for terrorism content in our Code at this time, but we will continue to build our evidence base in relation to this with a view to exploring such a recommendation in future.

14.140 In the meantime, we would welcome evidence from stakeholders that addresses our evidence gaps as listed above; particularly around accuracy and effectiveness, costs and impacts on users' rights.

## Other policy options for hash matching

---

### A broader recommendation for hash matching

14.141 We have also considered whether it might be appropriate to recommend that services use hash matching to detect other illegal content (beyond CSAM or terrorism content). We note that hashing can be used to tackle a range of illegal content. These may include for example, intimate image abuse, certain types of fraud, and potentially extreme pornography.

14.142 However, we currently have limited information on how services can and do use hash matching to address these harms, and the likely effectiveness and cost of applying this technology to such harms, and are therefore not proposing to recommend the application of hash matching technologies to other forms of illegal content in our codes of practice at this

time. We would welcome evidence in response to this consultation on the application of hash matching technologies to these and other kinds of illegal content.

14.143 We will continue to build our evidence base in this area with a view to potentially revisiting such a recommendation in future.

## URL detection

---

### Introduction

14.144 Content on a service which includes links to illegal content present on other internet services has the potential to cause serious harm to users of the service or other individuals. These links could take the form of hyperlinks, or the text of a web address. This section describes how the risk of harm posed by such links can be mitigated by detecting and removing links to illegal content using URL lists.

14.145 In this context, URL lists are lists of URLs at which illegal content is present. Some services compile their own URL lists, but often such lists are maintained by third parties such as NGOs, law enforcement bodies, or providers of safety tools. This section concerns the use of URL lists by U2U services; the use of URL lists by search services to de-index search content that is illegal content is discussed in Chapter 15 (Automated Search Moderation).

14.146 U2U services can compare user-generated content on their service to URL lists to detect links to illegal content. This comparison could take place at the point a user generates the content, or at a later point (for instance by scanning content present on the service). The service can then take down or obscure detected links so that users are unable to view the link or access the URL via that link.<sup>232</sup>

14.147 In this section, we consider recommending the use of URL detection technology to detect certain URLs shared on U2U services. This can involve 'direct matching' or 'fuzzy matching' (including 'case invariant' matching). The former will only return a match if a link shared by a user exactly matches a URL on the URL list. The latter can return a match despite changes made to the text.

14.148 URL lists can also vary in their approach, and may include either specific webpages or in some cases list at the domain level (e.g., a whole website).

14.149 Broadly, there are two ways in which services can obtain a URL list:

- ) They can set up and maintain their own internal URL list, which may involve drawing on URLs contained within third party provider URL lists; or
- a) They can procure a list from a third party provider, who may or may not also provide the detection technology or carry out detection for the service.

## CSAM URL detection

---

14.150 We have considered the case for recommending that services use URL detection to detect and remove URLs known to host CSAM. This section explains our considerations behind provisionally recommending certain services use URL detection technology effectively in

---

<sup>232</sup> Users may also be prevented from accessing a URL as the result of the operation of network-level filtering by their internet service provider or as a result of filters applied at a local level (such as parental control software or an organisation's IT policies).

relation to content communicated publicly by means of the service to detect URLs at which CSAM is present, or which include a domain which is entirely or predominantly dedicated to CSAM, and swiftly take these down (or prevent them from being posted).

## Harms that the measure seeks to address

- 14.151 As explained in paragraph 14.31 above, 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 CSAM, or which advertises the distribution or showing of CSAM.
- 14.152 References to CSAM in this section include CSAM in the form of images, video, written or audio content.
- 14.153 Links to CSAM (including to hosting locations where CSAM is stored) are shared widely across the open web. CSAM links are shared across a range of services, though some service types are particularly high risk.
- 14.154 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 or store content on their device. In addition, evidence suggests that perpetrators are now sharing CSAM links for financial gain.<sup>233</sup>
- 14.155 CSAM link sharing is considered a significant and growing concern by those across government, NGOs and industry.<sup>234</sup> WeProtect’s 2021 Global Threat Assessment highlighted the growth of ‘on-demand’ access to CSAM as opposed to the curation of personal collections, often through accessing links posted across multiple sites that lead to file hosts.<sup>235</sup> 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.<sup>236</sup>
- 14.156 URL detection would seek to reduce the sharing of links to CSAM and so mitigate the harms this causes to users and other individuals. The harms caused by the sharing of links to CSAM are essentially the same as described in the proposed hash-matching measure.<sup>237</sup> These harms are explored above at paragraph 14.38 and are discussed in more detail in the Register of Risks.<sup>238</sup>
- 14.157 In addition to harms associated with CSAM generally, whether shared via links or image-based CSAM, there are some specific harms which have greater relevance for CSAM URLs than image-based CSAM.
- 14.158 First, CSAM URLs may not be immediately recognisable as links to CSAM, or may be falsely advertised as being links to non-illegal content. As a result, it is more likely that a user may

---

<sup>233</sup> IWF, 2022. [Criminals blatantly spam links to child sexual abuse material online](#). [accessed 15 June 2023].

<sup>234</sup> Meta (Davis, A.), 2020. [Facebook Joins Industry Effort to Fight Child Exploitation Online](#). [accessed 14 June 2023]; [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 15 June 2023].

<sup>235</sup> WeProtect, 2021. [Global Threat Assessment 2021](#). [accessed 10 June 2023].

<sup>236</sup> Gov.UK, 2023. [Safety Tech Challenge: link sharing of Child Sexual Abuse Material](#). [accessed 10 June 2023].

<sup>237</sup> 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.

<sup>238</sup> See Volume 2: Chapter 6C CSEA (grooming and CSAM).

click on a link not realising it is CSAM, and inadvertently view CSAM. The impact of inadvertent viewing of CSAM is set out above, but of particular note is that accidental viewing can lead to more regular viewing; one study found that 51% of regular CSAM viewers first viewed CSAM accidentally.<sup>239</sup>

14.159 Second, as is described in the Register of Risk, a recent development in online CSAM is the rise of Invite Child Abuse Pyramid (ICAP) sites. Evidence indicates that a high volume of links to these sites are being shared across services to generate traffic and, ultimately, revenue for the owners of these sites. In addition to resulting in increased inadvertent viewing, the revenue generated by this form of link sharing is likely to be used to perpetrate further harm, including child sexual offences.<sup>240</sup>

## Options

14.160 We have considered whether to include in our CSEA Code of Practice a recommendation that services use technology to proactively identify CSAM URLs.

14.161 As with the hash matching measure, we have considered what an effective URL detection measure could look like. This discussion is set out at Annex 15, which considers each of the following points:

- a) The type of technology that should be used (for example, direct or fuzzy matching);
- b) The URL list used by services (including how it is maintained and the granularity of URLs included in the list);
- c) The breadth of content that is scanned (and when) on the service (e.g., new content only or including all existing content); and
- d) The degree of human review required over content identified by the technology.

## Outline measure

14.162 In light of that discussion, our proposed measure includes the following features:

- a) The use of a form of URL detection technology to detect direct matches in user-generated content to a list of known CSAM URLs. The technology would analyse written user-generated content communicated publicly by means of the service;
- b) The service should 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);
- a) Content detected to be a CSAM URL should be swiftly taken down (or prevented from being generated, uploaded or shared);
- b) The list of known CSAM URLs should be sourced from a person with expertise in the identification of CSAM, and arrangements should be in place to identify CSAM URLs, and to ensure the accuracy of the list (including when adding URLs to the list, and by reviewing the list to remove URLs where appropriate (e.g. where CSAM at the URL has been removed). The list (and any copy held for the purpose of the measure) should also be secured against security compromises through attacks by bad actors;

---

<sup>239</sup> CSAM Users in the Dark Web: Protecting Children Through Prevention, 2021.

<sup>240</sup> IWF, 2022. [The Annual Report 2022](#). [accessed 25 May 2023].

c) The list should be regularly updated, and the service should compare content to the latest version of the list.

14.163 We consider that in most instances, it would be appropriate for the URL included on the list to be the URL of the specific webpage at which CSAM is hosted (rather than listing whole domains), to avoid ‘over-blocking’ of legitimate content. (In this regard, we emphasise that CSAM is not limited to indecent and prohibited images: as explained above, such URLs would include material that contains advice about grooming or abusing a child sexually, or obscene articles encouraging the commission of CSEA offences. Importantly, they would also include URLs which include content linking or otherwise directing users to CSAM, or advertising 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).) In addition, however, we also consider that it would be appropriate to list at domain level where the domain is entirely or predominantly dedicated to CSAM.<sup>241</sup> This is likely to be more effective and efficient than listing each individual URL containing CSAM, given that these may alter frequently. However, we would expect services to ensure that the provider of the URL list has arrangements in place to ensure that listing at domain level only occurs in such cases.

14.164 Having provisionally identified what we think an effective URL detection technology recommendation could in principle look like, we then consider each of the following:

- the degree of accuracy, effectiveness and risk of bias from technology deployed in accordance with the outlined option;
- the extent of any interference with users’ rights to freedom of expression and privacy from such an option; and
- the costs of such an option.

14.165 We then consider for which services this might be proportionate, and set out our proposed recommendation.

## **Accuracy, effectiveness and lack of bias**

14.166 URL detection is a content identification technology, and so is regarded as a use of proactive technology for the purposes of the Act. As such, Ofcom must have regard to the degree of accuracy, effectiveness and lack of bias achieved by the technology in deciding whether to recommend URL detection. This is determined by an assessment of the technology itself and the URL list.

14.167 In respect of accuracy, deploying a form of URL matching technology to detect direct matches with URLs on a list should be highly accurate. Direct matching is a well-established, well-understood and straightforward mechanism for detecting text content on services. Whether matched content is a CSAM URL will depend on the accuracy of the URL list. The option we have outlined has elements designed to ensure that CSAM URLs are accurately included on the list, including that lists should be sourced from a person with expertise in the identification of CSAM and that arrangements should be in place to secure that CSAM URLs are correctly identified before being added to the list, and reviewed and removed if

---

<sup>241</sup> See further paragraphs A15.73 to 77 of Annex 15.

they are no longer CSAM URLs. We consider these will substantially mitigate the risk of content being incorrectly identified as a CSAM URL.

- 14.168 We consider direct matching to be highly effective in detecting direct matches to known CSAM URLs on a CSAM URL list.
- 14.169 Effectiveness also depends on the completeness, accuracy, and regular deployment of the URL list being used. The option we have outlined includes several elements designed to ensure the URL supports the effectiveness of the measure, such as ensuring that there are arrangements to identify suspected CSAM URLs and to regularly update the list, and that services compare content to the latest version available of the list.
- 14.170 The main risks of bias occur with the compilation of the URL list, rather than the technology. For example, addition of URLs to the list depends on where online the content is, how it is detected (e.g. through AI machine learning models, web crawling, or human analysts), and the assessment of content as CSAM (e.g. 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 and belonging to minority groups, but these are mitigated by the elements which promote the accuracy and effectiveness of the list.
- 14.171 We also consider it unlikely that the technology could be attacked by bad actors to increase the risk of false positives being generated, although we recognise that the URL list itself may be vulnerable to security compromises through attacks by bad actors. Perpetrators may attempt to attack services to make measures less effective. For example, the simpler the implementation of the technology, the higher risk this represents that the service can be attacked by bad actors to gain access to the URLs in question. As such, the measure includes that URL lists should be secured from security compromises through attacks by bad actors, and that appropriate measures should be taken to secure any copy of the list held by or for the service.
- 14.172 Overall, we consider that the use of this proactive technology will be effective in reducing the dissemination of CSAM by targeting a key technique that perpetrators use to share this content through a specific functionality. Deployment of URL lists would therefore deliver very significant benefits in mitigating against the harms caused by the availability of CSAM online. This includes reducing the harm caused by sharing of CSAM to victims and survivors, as well as reducing intentional viewing of, or unintentional exposure to, this content and subsequent contact sexual abuse. We have set the benefits of reducing availability of CSAM out in more detail in our discussion of hash matching above.

## Costs and risks

- 14.173 Costs will comprise both one-off costs of developing and implementing the URL detection tool and ongoing costs of maintaining the system and the required software, hardware and data. For the purposes of assessing costs, we assume a service is not undertaking a complementary measure, though costs set out are likely to be less for services already undertaking such measures.
- 14.174 Table 14.2 below summarises the costs detailed in this subsection for small services (low cost estimate) and large services (high cost estimate).

**Table 14.2: One-off and on-going costs associated with implementing a URL detection tool**

Lower estimate		Upper estimate	
Build cost (one-off)	Maintenance costs (ongoing, annual)	Build cost (one-off)	Maintenance costs (ongoing, annual)
£20,000	£5,000	£280,000	£70,000

## One-off costs

- 14.175 The one-off cost of developing the URL detection tool will primarily come in the form of labour costs. The key skillset required will be software engineering, though there may also be involvement from other skillsets (e.g. project management, legal, etc.). We understand that URL detection is complementary to other measures that a service may already be implementing (for example, keyword detection). Services which are already implementing a complementary measure would likely be able to do URL detection with a small additional cost, primarily in the form of software engineering time.
- 14.176 We understand that direct matching presents lower implementation costs than other forms of URL detection, such as a fuzzy matching system. Through our engagement with industry experts, we understand that this measure will be more straightforward for smaller and simpler services to implement, while larger and more complex services may require additional resource.
- 14.177 Large services are more likely to have more complex operational structures and service changes may involve input from more professions. The complexity of the service itself will also impact on costs as changes will require more resource where a service has multiple products that require integration. Based on this, we estimate it may take small and medium sized services approximately 2 months of full-time work for a software engineer to undertake the initial set-up of a CSAM URL detection system. For large services, this may increase to the equivalent of 16 months. Based on engagement with industry experts, we estimate this will also require a similar amount of time input from a combination of other professionals. Details of our assumptions on salaries are included in Annex 14. We estimate that these one-off product development costs may range from £20,000 to £280,000. We expect the lower estimate to be more reflective of smaller and less complex services and the upper estimate to be more reflective of larger and more complex services.

## Ongoing costs

- 14.178 There will be ongoing labour costs associated with maintaining and updating the system, such as integrating a new URL list. This will primarily require software engineering skills. To ensure we are fully capturing potential costs, we assume again that this requires a similar time input from software engineers and other skilled professionals. In line with our standard cost assumptions set out in Annex 14, we assume this to be approximately 25% of the initial set-up costs, ranging from approximately £5,000 to £70,000 annually.



- 14.179 In addition, services implementing this measure would need to use a URL list supplied by a third party, which is likely to have costs associated such as membership fees.<sup>242</sup>
- 14.180 We do not anticipate that this recommendation will incur additional material human moderation costs for the service; because the recommendation is for direct matching, we view the risk of false positives to be low and anticipate that any potential additional human moderation could be managed by existing teams.
- 14.181 We are not aware of technical difficulties or barriers that may be in place for services ingesting URL lists.
- 14.182 We note that some implementations of the URL list may cause particular security concerns and be vulnerable to security compromises through attacks by bad actors, especially those using simpler models. As such, services would need robust cybersecurity measures to protect against this risk and we have recommended this as part of this measure. Services may incur additional costs through implementation of security measures, though we note that many services are likely to have these security protections in any case, as a normal cost of operating their services.
- 14.183 We also consider it appropriate for services to put in place protocols for those dealing with the URL list to ensure that the material is handled safely. This will include some costs. Actions taken by services could include, for example, background checks on staff members, as well as ensuring that only named individuals are permitted access and involved with the implementation, testing, or review of the URL detection technology.
- 14.184 Services may choose to commission a third party to operate the URL detection and removal system, rather than undertaking this work internally. While existing services are currently incurring these costs, we anticipate that other providers may create URL lists as the market for such safety tools develops. This means costs are incurred by those providers, as well as for services using those providers depending how their fee structure is managed.
- 14.185 As outlined above, our evidence on costs suggests that the measure could be less expensive for smaller and less complex services than larger services.

### Dynamic effects on the market

- 14.186 Similar to our considerations for hash matching, there could be challenges with the capacity of current database providers to support a significantly more widespread use of URL detection in a short time frame.
- 14.187 We recognise that there may however be benefits for the market should there be a wider deployment of URL detection technology. In particular, an increase in the number of internet services that implement URL detection to detect known CSAM would likely have a positive impact on the level of interest, investment, and innovation in the use of URL detection to reduce known CSAM online. A welcome impact of this would be a likely reduction in the cost of adopting URL detection technology and an improvement in the effectiveness of this.

---

<sup>242</sup> To provide an example, the IWF currently provide a URL list, in addition to other services including an image hash database. At the time of writing (June 2023), the IWF's membership fees to support their work range from £1,000 to over £80,000 for the largest services annually, based on industry sector and company size. The membership list available online demonstrates that smaller services are accessing this membership at the lower end of this cost. Source: IWF, 2023. [Our Members](#). [accessed 15 June 2023].

## Rights impacts

- 14.188 This subsection discusses the potential impacts of the CSAM URL detection option we have outlined on individuals' rights – in particular, to privacy under Article 8 ECHR and to freedom of expression under Article 10 ECHR.
- 14.189 As described in relation to the proposed hash matching measure, interference with qualified rights may be justified on specified grounds such as the prevention of crime, the protection of health and morals, and the protection of the rights of others.
- 14.190 The measure seeks to address harms associated with the dissemination of CSAM through posting links on relevant services. These harms are essentially the same as described in the proposed hash matching measure (with the exception being that the CSAM concerned is not limited to images and videos, and could include text or audio content). The analysis in relation to the hash matching measure described the public interest that exists in measures which aim to 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), and that public interest applies equally to measures which focus on disrupting the means by which CSAM is accessed, as to measures which focus on taking down CSAM.

## Users' freedom of expression

- 14.191 Our provisional assessment is that any impact on users' rights to freedom of expression should be slight.
- 14.192 So far as links to CSAM are correctly detected and taken down, such content either does not engage Article 10 ECHR at all or otherwise restrictions in relation to that content are clearly justified to protect overriding public interests. In this regard, we note that links to CSAM will often themselves be 'priority illegal content' for the purposes of the Act (such as where they encourage or assist the commission of an offence committed by accessing an indecent image of a child). We recognise that there are cases where links are shared by those who do not have a sexual interest in children, including in outrage or disgust, but such sharing can still cause serious harm.
- 14.193 We consider that there should be few cases where links are incorrectly taken down. URL matching technology used in accordance with the option we have outlined should be highly accurate, given that it recommends direct matching to a list of URLs sourced from a third party with expertise in the identification for CSAM and makes further provision to ensure the accuracy of the list and its use, as follows:
- 14.194 the need for arrangements to secure that CSAM URLs are correctly identified before being added to the list, and to review CSAM URLs on the list and remove them where appropriate (for example because the CSAM present at the URL has been taken down);
- 14.195 the need for the list to be regularly updated and for the service to use the latest available version; and
- 14.196 for both the list and any copy of the list held for the purposes of the service to be secured from unauthorised interference (which would safeguard the list from the risk of bad actors adding URLs to the list for malicious purposes).
- 14.197 We acknowledge however that there could be cases where an URL has been incorrectly included on the URL list as a CSAM URL, or where a URL continues to be blocked for a period

after CSAM has been removed from it. In these cases, if the affected user complains, services could refer the complaint to the third party from whom the list has been sourced to review whether the URL should be removed from the list.

- 14.198 We also recognise that there may be some interference with freedom of expression insofar as the content present at a URL includes legitimate content as well as CSAM (including where a link to that URL is taken down, but the URL is subsequently removed from the URL list once the CSAM present at the URL is taken down).
- 14.199 The option we have outlined also provides for URLs to 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, but consider that this is justified to protect public interests (given the risk that users accessing such URLs will go on to encounter CSAM).

## Users' privacy

- 14.200 Consistent with the Act's constraints on proactive technology, the option only applies detection of CSAM URLs within content that is communicated publicly by means of the service. As such, it should not affect the confidentiality of communications.
- 14.201 Any processing of personal data for the purposes of the measure should be limited to the automated analysis of the relevant content to detect whether it consists of or includes a URL, and is unlikely to engage users' rights to privacy under Article 8 ECHR.
- 14.202 Insofar as links to CSAM URLs are themselves considered to be CSEA content, service providers may be required to report the links, and details of the user responsible for posting the link, to the Designated Reporting Body housed in the NCA in accordance with section 66 of the Act (once brought into force, and as further specified in regulations made under section 67).
- 14.203 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. In cases where a link has been incorrectly detected as a CSAM URL, though, this reporting would constitute an interference with the affected users' privacy. However, as explained in relation to the hash matching measure, there will be processes in place to ensure investigatory action is only taken in appropriate circumstances.

## Provisional conclusion

- 14.204 As set out at paragraphs 14.152 – 14.154 the posting and sending of CSAM URLs causes very significant harm.
- 14.205 Our provisional view is that analysing user-generated content to detect matches with known CSAM URLs can be an effective means of content moderation to enable services to proactively identify and tackle the dissemination of CSAM. As set out at paragraphs A15.63 – A15.77 of Annex 15, our provisional view is contingent on services deploying an appropriately maintained list of CSAM URLs and having sufficient oversight to ensure the process is working as intended.
- 14.206 We recognise that recommending CSAM URL detection in Codes could result in material costs for some services, and could result in some interference with user rights. We consider that the design of our proposed measure, and in particular the safeguards for the protection

of users' freedom of expression and privacy incorporated within it, mean that any potential impacts on those matters are limited. Our view is that any such impacts are justified by the substantial public interest in the prevention of crime, the protection of health and morals, and the protection of the rights of victims and children that the measure is designed to achieve, and are proportionate to the anticipated benefits from disrupting the dissemination of CSAM through posting links. We also do not consider that there is a less intrusive way of achieving these aims.

- 14.207 In addition, the severity of the harm we would be addressing and the benefits that would result from diminishing this, are sufficiently great that we provisionally consider it would be proportionate to recommend that at least some services deploy URL detection for known CSAM URLs. We are also aware that where a service is already undertaking a complementary measure (for example, we understand there could be synergies between this measure and hash matching), the costs associated with implementing URL detection could be less than those set out at Paragraph 14.170. We consider that our proposed measure (which is discussed in more detail at Annex 15, and set out in draft in Annex 7) could help prevent the dissemination of hundreds of thousands of URLs identified as containing CSAM each year.
- 14.208 There is, however, a question as to which services we should recommend be in scope of this measure. As described previously, to inform our view we have had regard to the principle that measures should be proportionate, taking into account the risk of harm presented by services and the size and capacity of providers of services.
- 14.209 When considering which U2U services it would be proportionate to recommend a URL detection measure to, we considered the same options as set out in the hash-matching measure above (i.e. (1) apply the measure to all U2U services, (2) apply the measure to all services at least at low risk of CSAM URLs in their risk assessment, (3) apply the measure to all services at medium or high risk of CSAM URLs in their risk assessment, or (4) apply the measure to large services at medium or high risk of CSAM URLs in their risk assessment and other services at high risk of CSAM URLs in their risk assessment and that have more than 700,000 monthly UK users). Broadly the same considerations apply to this assessment as to our assessment of which services to propose hash matching apply to. However, unlike in the case of image-based CSAM, the evidence is not clear that file-storage and file-sharing services generally pose a higher risk for the dissemination of CSAM URLs than other types of service.
- 14.210 Taking account of the risk to users and other individuals, and the severity and nature of the harm associated with its dissemination, as well as to the different size and capacity of services, our provisional view is that it is proportionate to recommend the use of URL detection for CSAM URLs in our CSEA Code of Practice in relation to the following services:
- large services (i.e. that have more than 7 million monthly UK users) that are at medium or high risk for CSAM URLs;<sup>243</sup> and

---

<sup>243</sup> A service is likely to be high risk for CSAM URLs if it allows text or hyperlinks to be posted or sent; and any of the following applies: i) the service has systematically been used by offenders to post or send CSAM URLs; ii) the service allows users to post or send content without creating an account. A service is likely to be medium risk for CSAM URLs if the service allows text or hyperlinks to be posted or sent and does not meet the criteria for high risk; and any of the following applies: i) there is evidence that the service has been recently used by offenders to post or send CSAM URLs; ii) the service has two or more relevant risk factors associated with CSAM URLs in Ofcom's Risk Profiles, in addition to allowing text or hyperlinks to be posted or sent. Please refer to the Risk Assessment Guidance for more information.

- other services that are at high risk of CSAM URLs and have more than 700,000 monthly UK users.

## Terrorism content URL detection

---

### Harms or risks that URL detection seeks to address

- 14.211 As described above, terrorism content is disseminated across an eco-system. Experts emphasise how terrorist groups have increasingly adopted a multiplatform approach for content dissemination. Similarly, from our own work, we understand how footage of far-right shootings are disseminated across a plethora of services.<sup>244</sup>
- 14.212 There is evidence that supporters of Islamic State have tried to circumvent content blocking technology on larger, well-resourced sites by ‘outlinking’ to smaller platforms with limited resources.
- 14.213 The site hosting URLs, or outlinks, is sometimes referred to as a ‘beacon’ platform. From the beacon platform these outlinks direct users to file-storage and file-sharing services, terrorist operated websites, and social media platforms on which the terrorism content is hosted.<sup>245</sup>
- 14.214 URL links are a key method by which users can be transported to terrorism content. A research study into Islamic State bots understood how they operate one of three key functions; publishing content, moderating discussions, and acting as gatekeepers. The gatekeeping role involves the dissemination of URL links.<sup>246</sup>
- 14.215 We note that these URLs can be generated at pace and at scale, including using mirroring services which generate banks of URLs for dissemination of terrorism content.
- 14.216 Services are aware of the dangers posed by outlinking towards terrorism content. For example, YouTube’s policies in relation to terrorism content refer to URLs.<sup>247</sup> We also understand that the ways that these URLs are being used is also evolving. For example, Professor Stuart Macdonald conveyed to us how URLs used to be used as ‘throwaway’ items by Jihadists, but more recently he is witnessing some URLs being used repeatedly.<sup>248</sup>
- 14.217 The scale and pace of dissemination of terrorism content via URLs present a high risk of harm, which can result in serious real-world consequences and ultimately a risk to life.

### Options

- 14.218 As outlined above, we understand that URLs including terrorism content can cause significant harms to citizens and consumers in the UK. We consider that measures applied to reduce the dissemination of these URLs would have a positive impact on harm both online and in the real world.

---

<sup>244</sup> Ofcom, 2022. [The Buffalo Attack: Implications for Online Safety](#). [accessed 10 May 2022].

<sup>245</sup> Hall, J. and Macdonald, S., 2023. [Online Safety Bill: Distinguishing between public and private communication](#). [accessed 12 April 2022].

<sup>246</sup> Alrhoun, A., Winter, C. and Kertész, J., 2023. [Automating Terror: The Role and Impact of Telegram Bots in the Islamic State’s Online Ecosystem](#). [accessed 13 April 2023].

<sup>247</sup> “Please note that these policies also apply to external links in your content. This can include clickable URLs, verbally directing users to other sites in video, as well as other forms.” YouTube, 2023. [YouTube Help](#). [accessed 20 June 2023].

<sup>248</sup> From meeting between Ofcom and CYTREC on 29/05/23

14.219 We understand that there are some organisations that are taking steps to reduce the dissemination of terrorism URLs, such as:

- Tech against Terrorism employ a database known as the Terrorist Content Analytics Platform (TCAP) that alerts platforms to violative URLs, in order to prevent the further dissemination of content. This database of URLs is substantial and in the two and half years since November 2020 it alerted over 100 different tech platforms to over 25,000 pieces of terrorist content.<sup>249</sup>
- The GIFCT 2021 Taxonomy Report recommended broadening the scope of content covered to include URLs.<sup>250</sup> The GIFCT is now looking to integrate hashes of URLs flagged by the TCAP that correspond to content produced by entities on the UN Consolidated Sanctions List and content that activates GIFCT’s Content Incident Protocol into its hash sharing database.
- [CONFIDENTIAL ✕].<sup>251</sup>

14.220 We also understand that, alongside ‘beacon platforms’, file-storage and file-sharing sites play an important role in caching and storing terrorism content. As noted by Tech against Terrorism’s insights publication of April 2023, file-storage and file-sharing sites “were by far the most exploited platform type in terms of volume of content, with 28,724 URLs submitted to the TCAP, representing 72% of all submissions.”<sup>252</sup>

14.221 However, at this stage, we consider we require further evidence in order to propose a recommended measure tackling the harm created by the dissemination of these links, in particular about the following areas:

- the utilisation of different approaches to URL detection and blocking technology to prevent the dissemination of these links (e.g. which type of service to target a measure at, and how best to apply it);
- the availability and accuracy of third party lists (including which providers can offer these services, any accessibility constraints, and how accurate these lists are);
- the extent of any potential interference with users’ freedom of expression presented by URL detection for terrorism content, and the steps services could take to address these; and
- the potential costs associated with implementing an approach, including setup and maintenance.

## Provisional conclusion

14.222 Sharing terrorism content via URLs is an evidenced way that bad actors are disseminating a wide variety of terrorism content, resulting in serious harm both online and in the real world.

14.223 We are aware of some stakeholders working to reduce the spread of terrorism content via URLs. We do consider that some form of URL detection is likely to be an effective measure to

---

<sup>249</sup> Tech against Terrorism, 2023. [Patterns of Online Terrorist Exploitation](#). [accessed 10 May 2023].

<sup>250</sup> GIFCT, 2021. [Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps](#). [accessed 18 April 2023].

<sup>251</sup> [CONFIDENTIAL ✕].

<sup>252</sup> Tech against Terrorism, Patterns of Online Terrorist Exploitation, 2023.

tackle the dissemination of terrorism content across the ecosystem. However, as noted above, we consider that we need further evidence to clearly articulate what measure we would recommend in our terrorism Code of Practice, including its cost implications and potential rights impacts.

- 14.224 Therefore, we are not proposing to include a recommendation relating to the use of URL detection technology for the purposes of detecting terrorism content in our first Code of Practice. However, as explained at the start of this chapter, this should not be read as an indication that we consider taking such measures to be ineffective or disproportionate or an indication that services should stop using this technology. We welcome innovation and investment in safety technologies such as URL detection technology, and plan to consider this further for future versions of our Codes. We therefore welcome stakeholders' views (and evidence) on the use of this technology to reduce the dissemination of terrorism content, and we will continue to build our evidence base in relation to this with a view to exploring such a recommendation in future.

## Keyword detection

---

### Introduction

- 14.225 Keyword detection is an established technique used in information retrieval and data analysis, which involves searching for specific text (such as words, phrases and special characters) known as 'keywords' within a set of data sources. The process entails matching the selected keywords against the content of the data, and then presenting the results that contain exact or related matches to the provided keywords.
- 14.226 Keyword detection plays a vital role in text-based content moderation in detecting, monitoring and filtering violative and illegal content. It can be utilised to identify potentially illegal or violative content to be removed automatically or, for example, flagged for human review.
- 14.227 There are different ways of deploying keyword detection technology. It can be deployed as a standalone technique to search and detect content relying on provided keywords, and we refer to this as 'standard' keyword detection technology in this consultation. It can also be used in combination with machine learning techniques – in particular, keyword lists and content flagged by keyword detection can be used to help train machine learning models or be used in combination with machine learning models to differentiate between illegal and legal content on a service. Similar to our discussion of URL detection above, we understand that there are two main types of standard keyword detection:<sup>253</sup> direct matching, which requires words to exactly match those on the keyword list, and fuzzy matching (including case invariant matching), which allows for partial similarities, providing a degree of tolerance to variations in the data.
- 14.228 In this section, we outline and assess the case for our policy proposals for standard keyword detection technology in relation to U2U services.
- 14.229 We understand that more advanced keyword search, detection, or filtering methods may already be in use by some services with some of them leveraging ML/AI or hybrid/layered approaches with a combination of both. However, given the limited evidence that we have

---

<sup>253</sup> NetClean, 2023. [Technologies to stop CSAM: Keyword Matching](#) [accessed 14 June 2023].



at present about the accuracy of such technologies and the potential for them to generate higher volumes of false positives, we are not proposing to recommend services use these in our Code of Practice at this time. Nevertheless, where services have implemented standard keyword detection technology in accordance with our proposed measure, our measure provides flexibility to services should they wish to deploy ML/AI on identified content before deciding whether to take it down. We also welcome innovation in these areas and recognise that some services may choose to use more advanced keyword detection technology in combination with ML/AI to identify potentially illegal content (and comply with their illegal content safety duties) as an alternative measure to adopting the measures set out in our Codes of Practice.

## Keyword detection regarding articles for use in frauds

---

14.230 We have considered the case for recommending that U2U services use standard keyword detection technology to effectively detect content which is likely to amount to a priority offence regarding articles for use in frauds, and then consider such content in accordance with their internal content moderation policy.<sup>254</sup>

### Harms that the measure seeks to address

14.231 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. Content amounting to any of these offences is therefore recognised by Schedule 7 of the Act as priority illegal content, and we refer to it in this section as content amounting to an offence concerning articles for use in frauds.<sup>255</sup>

14.232 Articles for use in frauds can include stolen bank card details (sometimes referred to as 'Fullz') and personal identifiable information, as well as 'Fraud bibles' and instruction manuals providing guidance on how to carry out fraudulent activity.

14.233 Research completed by Ofcom<sup>256</sup> as well as research by other organisations<sup>257</sup> has shown that some in-scope services (including social media services and search services) are being used by fraudsters to supply, or offer to supply, articles for use in frauds.<sup>258</sup> Our research suggests that, whilst it is unlikely to be encountered accidentally by internet users, this type of content is often very discoverable by criminals and likely to be prevalent on the open web

---

<sup>254</sup> We discuss in Chapter 12 our proposed recommendations regarding content moderation on user-to-user services in general, including that such services have in place content moderation systems or processes designed to take down illegal content swiftly, and that all large services and those that have assessed themselves as having a medium or high risk for any type of offence should set internal content policies which specify how content moderation systems and processes moderate content and resource them accordingly.

<sup>255</sup> As explained in Ofcom's draft Illegal Content Judgment Guidance, content online is most likely to be 'offering to supply' articles for use in frauds.

<sup>256</sup> Ofcom, 2023. [Online Content for use in the commission of fraud -accessibility via search services. 18 September 2023](#) [accessed 18 September 2023].

<sup>257</sup> Okpattah, K., 2021. [Social Media fraud: The influencers promoting criminal scams](#), BBC, 16 August 2021. [accessed 20 May 2023].

Cifas and Forensic Pathways, 2018. [Where do fraudsters hunt for data online?](#), 19 June 2018 [accessed 26 September 2023].

Lipson, F., 2020. [Your life for sale: stolen bank details and fake passports advertised on social media. Which?](#) 30 April 2020. [accessed 13 September 2023].

<sup>258</sup> SEON, 2023. [Fullz](#). [accessed 20 May 2023]; Fraud.net, 2023. [Fullz](#). [accessed 20 May 2023].

and dark web – often on online forums.<sup>259</sup> Research commissioned by CIFAS in 2018 revealed that ‘Fullz’ packages including personal data and financial information sell for about £31 on the surface web, while data held on the magnetic strip of bank cards sells for around £70.<sup>260</sup>

- 14.234 In particular, a defining characteristic of this type of content is the dense combining of key terms (e.g., ‘Fullz’ [CONFIDENTIAL X]) often alongside the use of multiple fraud-related hashtags (e.g. [CONFIDENTIAL X]). Our research suggests that very specific keywords tend to be used to indicate articles for use in frauds, and that – particularly when combined - these are unlikely to be used in any legitimate context, other than academic and news articles discussing the offence of making or supplying articles for use in frauds. Fraudsters use this kind of explicit and unguarded language in order to maximise the discoverability of their posts by persons seeking to engage in fraudulent activity. Notably, an investigation by *Which?* identified several profiles, pages and groups across multiple social media services by “*searching just a few slang terms used by fraudsters*”. These profiles, pages and groups “*advertised a mixture of stolen identities, credit card details, compromised Netflix and Uber Eats accounts, as well as fraud 'how to' guides and even fake passports made to order*”.<sup>261</sup>
- 14.235 Offences concerning articles for use in fraud are likely to arise chiefly in the communication of two or more potential perpetrators of offences. A perpetrator makes use of the ability to openly search on these services to locate other perpetrators offering to supply articles for use in frauds. This suggests that content of this nature is less likely than other offences to come to a service’s attention through standard user reporting. The use of keyword detection technology can therefore be a means to bolster the service’s ability to detect content of this nature.
- 14.236 Whilst the making or supply of articles for use in frauds is a priority illegal offence by itself, it also facilitates and enables the commission of other priority illegal fraud offences, including high harm frauds such as identity theft. For example, once criminals have acquired access to a package of stolen financial credentials and related personal information, these will then typically be used to undertake a wide range of secondary fraud activities, either online or offline. These include card-related fraud (e.g., the fraudulent purchase of goods/services/subscriptions, making payments to ‘money mule’ accounts to launder the proceeds of crime), or impersonation/identity fraud (e.g., stealing someone’s identity to take over or set up new bank accounts, email accounts or social media profiles to support fraudulent loan applications).<sup>262</sup>
- 14.237 Over recent years, there has been a significant rise in identity fraud cases and the scale of unauthorised fraud losses.<sup>263</sup> UK Finance has flagged that remote purchase fraud, where a criminal uses stolen card details to buy something online, remains the biggest category of losses at £395.7 million in 2022. Fraud on lost and stolen cards increased by 30 per cent compared to 2021 to £100.2 million in 2022 and card ID theft, where a criminal opens or

---

<sup>259</sup> Bodker, A., Connolly, P., Sing, O., Hutchins, B., Townsley, M. and Drew, J., 2021. [Card-not-present fraud: using crime scripts to inform crime prevention initiatives](#). [accessed 15 May 2023].

<sup>260</sup> Ashford, W., 2018. [Surface web used in private data sales](#). [accessed 15 May 2023].

<sup>261</sup> Lipson, F., 2020. [Your life for sale: stolen bank details and fake passports advertised on social media](#). *Which?* 30 April 2020. [accessed 13 September 2023].

<sup>262</sup> SEON, 2023. Fullz. DATADOME, 2023. [What are fullz? How do fullz work?](#) [accessed 20 June 2023].

<sup>263</sup> CIFAS, 2023. [Identity fraud cases reach all-time high as cost-of-living crisis bites](#). [accessed 20 June 2023].

takes over a card account in someone else's name, almost doubled to £51.7 million.<sup>264</sup>

Whilst the supply (and offers to supply) of articles for use in frauds online may not be a contributing factor in each instance of wider fraud, our understanding is that it is a contributing factor in many cases, enabling criminals to commit further frauds and inflict additional harm on the public.

14.238 The implications of these wider priority illegal fraud offences can be severe, depending on the specific circumstances. They can result in physical, emotional and psychological harm, both during the event and afterwards. A victim of identity theft can also be left vulnerable to future instances of fraud due to the sale of their stolen credentials, which would compound any financial and emotional impact. It is relevant to note that the severity of these impacts is not always directly linked to the scale of the original financial loss – though in many instances this may be a critical factor. Fraud can result in financial harm to the user, and their family. Ofcom research on fraud overall found that a quarter (25%) of those who encountered an online scam or fraud lost money as a result.<sup>265</sup> In addition, a third (34%) of those who have encountered an online scam or fraud claim that the experience has had an immediate negative impact on their mental wellbeing. Those who lost money as a result of an online scam are more likely to have been negatively affected in both the short and long term.

14.239 A measure that enables services to identify and remove content that amounts to a priority offence regarding articles for use in frauds should reduce wider instances of fraud and therefore harm to individuals. Specifically, it would:

- a) make it harder for fraudsters to market the proceeds of criminal activity, and therefore diminish the attractiveness of those original illegal activities (i.e., the theft of personal details);
- b) make discoverability of this content more difficult, ultimately reducing the ability to commit fraud using these credentials or guidebooks;
- c) limit easy surface web access to sources of stolen financial credentials, which means that opportunistic fraudsters will be disincentivised; and
- d) protect individuals from becoming victims of fraud, resulting in less financial and emotional distress.

## Options

14.240 We have considered whether to include in our Code of Practice a recommendation that U2U services deploy standard keyword detection technology to proactively identify content that is likely to amount to an offence concerning articles for use in fraud.

14.241 Consistent with other sections in this chapter, we have first considered what an effective standard keyword detection option could in principle look like. This is set out in Annex 15, which considers each of the following:

- The type of technology that should be used (for example, direct or fuzzy matching);

---

<sup>264</sup> UK Finance, 2023. [Over £1.2 billion stolen through fraud in 2022, with nearly 80 per cent of app fraud cases starting online](#). [accessed 20 May 2023].

<sup>265</sup> Ofcom, 2023. [Online fraud and scams](#). [accessed 16 March 2023].

- The steps services should take to ensure that they have access to an appropriate keyword list (and that the list is appropriately maintained);
- The breadth of content that is scanned (and when) on the service (e.g., for new content or for all existing content);
- What provision should be made about the technical performance of the technology; and
- What steps services should take when the technology identifies potentially illegal content.

## Outline measure

14.242 In light of that discussion, we consider that an effective measure would comprise the following elements:

- a) The use of fuzzy keyword detection technology to analyse content in the form of written material or messages which is communicated publicly on the service. This includes both analysing content that is already present on the service (within a reasonable time), and content that is thereafter uploaded to the service (before or as soon as practicable after it can be encountered by other users).
- b) The use of a suitable keyword list. To be suitable, appropriate steps should be taken to ensure that it:
  - i) contains only keywords that could not reasonably be expected to be used on the relevant service (either on their own or in combination with other keywords on the keyword list) except in relation to the commission of an offence concerning articles for use in frauds; and
  - ii) is sufficiently comprehensive.
- c) We have set out the minimum steps that we think should be taken in this regard, including the information that services should consider when compiling their keyword list, the use of appropriate measures to test the keyword list on a reasonable sample of content, securing the list against unauthorised access, interference or exploitation, and review of the keyword list at least every six months. Service providers would also need to ensure a written record is made about how they compile their keyword list, and about reviews and updates to the list.
- d) Ensuring that the technology is configured so that its performance strikes an appropriate balance between precision and recall, taking into account:
  - iii) the risk of harm relating to fraud, as identified in the service's latest illegal content risk assessment, and including information reasonably available to the provider about the prevalence of relevant content which amounts to an offence concerning articles for use in frauds on the service;
  - iv) the proportion of content detected by the technology that is a false positive;
  - v) the effectiveness of any systems and processes used to identify false positives before takedown; and
  - vi) reviewing this at least every six months (and at the same time as review of the keyword list). Service providers would also need to ensure a written record is made of how this balance has been struck.

- e) Ensuring that human moderators are used to review a reasonable sample of content detected by the technology within each review period. We have set out certain principles which they should take into account when deciding what is a reasonable sample. Service providers would need to ensure that a written record of the volume of content reviewed is kept, as well as information about how the principles have been taken into account when determining what is a reasonable sample.
- f) Ensuring that content detected by the technology is considered by the service in accordance with its internal content moderation policies.

14.243 Our provisional view is that this would be an effective measure which has sufficient clarity for providers while also providing an appropriate degree of flexibility as to how it is adopted.

14.244 Having provisionally identified what we think an effective keyword detection technology recommendation could in principle look like, we then consider each of the following:

- the degree of accuracy, effectiveness and risk of bias from technology deployed in accordance with the option we have outlined;
- the extent of any interference with users' rights to freedom of expression and privacy from such an option; and
- the costs of such an option.

14.245 We then consider for which services this might be proportionate, and set out our proposed recommendation.

## **Accuracy, effectiveness and lack of bias**

14.246 We understand that the keyword detection technology that we have considered is a relatively elementary form of content moderation which is established in a range of contexts.

14.247 Our analysis above and in Annex 15 suggests that the deployment of standard keyword detection technology could be an effective means of detecting content which amounts to an offence concerning articles for use in frauds online, particularly given that research carried out by Ofcom suggests that the terms associated with articles for use in frauds are often very specific.<sup>266</sup>

14.248 The accuracy and effectiveness of the technology will depend on both the content of the keyword list and the manner in which the keyword detection is conducted. The option we have outlined does provide flexibility to services as to the words included in their keyword list and the configuration of the keyword detection technology. However, the measure we have considered recognises this and includes principles to help ensure the accuracy and effectiveness of the technology. These include that:

- appropriate steps are taken by service providers to ensure that their keyword list only contains keywords which (either on their own or in combination with other keywords on the list) would not reasonably be expected to be used on the service except in relation to the commission of an offence concerning articles for use in frauds, and is sufficiently comprehensive; and

---

<sup>266</sup> Ofcom, 2023. [Online Content for use in the commission of fraud -accessibility via search services. 18 September 2023](#) [accessed 18 September 2023].

- the performance of the keyword detection technology be configured so as to strike an appropriate balance between precision and recall (with reviews of the performance of the technology at least every six months, including human review of a reasonable sample of detected content in each review period).

14.249 In light of the above, we would expect any content detected as a result of applying this measure to be highly likely to amount to an offence concerning articles for use in frauds. We recognise however that the keyword detection measure we are considering will enable services to identify content about which no prior illegal content judgment or determination has been made and that it may result in false positives. It may identify legitimate content (such as news articles or academic articles) which discuss the supply of articles for use in fraud. It is for this reason that we are not recommending that services take down all content detected by the technology, and are instead recommending that it be considered by services in accordance with their internal content moderation policies.

14.250 Whilst we are unable to predict how criminals might respond to this measure with certainty, we recognise that - as online services identify and remove this content via keyword detection - users may adopt new keywords (and combinations of keywords) to both conceal and highlight their activity. This could affect the accuracy and effectiveness of the measure over time, as well as its proportionality. However:

- the measure outlines that in-scope services take appropriate steps to ensure that the keyword list only contains keywords (and combinations of keywords) that would not reasonably be expected to be used on the service for any purpose other than in relation to articles for use in frauds, and is sufficiently comprehensive. We have also set out further detail on what those appropriate steps should include. Any changes in the words used to indicate or advertise articles for use in frauds should therefore be identified by services as part of their regular reviews of their keyword list, and we do not expect these changes should undermine the effectiveness of the option we have outlined.
- We recognise that the more often fraudulent actors change their language, the more often services will likely need to modify their systems to reflect these changes (and that this could impact on the proportionality of the measure under consideration). Similarly, we recognise the risk that criminals may seek in response to this measure to use more common words when advertising articles for use in frauds, and that this could impact the effectiveness and proportionality of the measure. However, we consider this to be a relatively low risk. We note that those who seek to supply articles for use in frauds online are incentivised to maximise the discoverability of their content for an audience looking to acquire such articles, and frequent changes in terms, or the use of more common words, would make discoverability and dissemination of this content more difficult. We would expect this to itself have positive impacts on citizens and consumers.

14.251 We do not consider that the measure we have detailed above (which does not include keyword detection based on machine learning/AI) should result in bias. The technology searches for fuzzy (and exact) matches in user-generated content based on a keyword list, and does not analyse or reach judgments based on user data or associated metadata more generally. We have also not seen evidence to suggest that the technology is biased (for example, based on users' protected characteristics).

## Costs and risks

14.252 From our engagement with industry experts, we understand that the main factors determining the cost of keyword detection technology include:

- the steps taken to compile the keyword list, including whether this is outsourced or in-house;
- how sophisticated the keyword detection technology is, including whether this is built in-house or outsourced;
- maintenance and review costs for both the keyword list and keyword detection technology; and
- the wider content moderation systems and processes into which the detected content is inputted.

14.253 We recognise that our understanding of costs is more limited in relation to this measure than other measures we are considering here (i.e., CSAM Hash Matching, CSAM URL detection). We also understand that establishing and running keyword detection technology, alongside human review for quality assurance purposes, may be costly. For this reason, we have considered a wide range of potential costs and are particularly interested in stakeholders' feedback, including evidence, on the likely costs associated with this option.

14.254 We are also mindful that we have less evidence about the use of keyword detection technology by smaller services, and that the cost of this option may represent a larger proportion of revenue for the smallest services. In particular, this may be the case for those services with low revenues who may experience proportionately higher costs (as not all costs associated with this option would scale in proportion to the size of the service).<sup>267</sup>

### One-off costs

14.255 Services that do not already have standard keyword detection technology in place would need to incur one-off costs to build this system. We would expect these costs to come in the form of labour costs.

14.256 The key skillset required will be software engineering, though there will likely also be involvement from other skillsets (e.g., project management, legal, etc.), particularly for medium and large services.<sup>268</sup> Larger services are more likely to have more complex operational structures and service changes may involve input from more professions. The complexity of the service itself will also impact on costs as changes will require more resource where a service has multiple products that require integration.

14.257 Based on a combination of engagement with industry experts and our own research and expertise, we estimate that it may take the equivalent of three months of software engineering time to develop and implement standard keyword detection technology, with similar resource required from other skillsets.<sup>269</sup> This represents a conservative estimate and

---

<sup>267</sup> For example, smaller services or services with low revenue may be less likely to have pre-existing specialist policy knowledge about offences concerning articles for use in frauds, or pre-existing engineering expertise to integrate keyword detection technology across their service.

<sup>268</sup> The exact combination of professions who may be involved in developing and implementing a fraud detection tool will vary significantly depending on the size and complexity of a service.

<sup>269</sup> For example, see: Sentropy Technologies, 2021. [How to build a content moderation system](#). [accessed 20 August 2023]



is more likely to be reflective of costs for smaller and less complex services. To reflect our currently limited evidence base on the costs of this measure for different kinds and sizes of services, we have also adopted an upper estimate for software engineering time of 18 months, which may be more representative of costs for the largest and most complex services. We estimate this will also require a similar amount of time input from a combination of other professionals. Our assumptions about salaries for different professions are included in Annex 14.

14.258 We therefore estimate the total one-off labour costs to set up the keyword detection technology may range from £30,000 - £320,000. Costs are likely to be greater for larger and more complex services where the technology may need to be implemented across multiple functionalities.

## Ongoing costs

14.259 Ongoing costs will include the following:

- a) Costs of compiling and review of the keyword list at least every six months;
- b) Costs of maintaining the keyword detection technology, and review of its technical performance at least every six months;
- c) Costs of human moderation of a reasonable sample of detected content; and
- d) Costs of considering detected content in accordance with the service's internal content moderation policies.

### *Keyword list*

14.260 Services that do not already have a keyword list in place will need to incur costs to do so, which will likely be greatest for the initial setup, though there will be costs associated with the ongoing review and updating of the keyword list. Services may choose to conduct their own internal research in the first instance (which we would expect to be a relatively low resource task)<sup>270</sup>, or engage a third party with relevant expertise. In either case, this option also sets out a number of additional steps that services will need to take to ensure their list is sufficiently accurate and effective, including consideration of previous content moderation decisions and testing of the list on the service.<sup>271</sup>

14.261 The costs of developing a keyword list are likely to be primarily labour costs<sup>272</sup>, although services may use external experts which could increase costs. We are aware that some services may choose to work with third party vendors, in combination with in-house efforts,

---

<sup>270</sup> Ofcom's internal research into the prevalence of articles for use in frauds on search services required the creation of a suitable shortlist of initial keywords. Initial desk research provided a range of starting terms from which the research team could build their initial list. The experience of this work suggests that this should be a relatively low resource task.

<sup>271</sup> These steps include that services have regard to the outcomes of reviews of content carried out by human moderators and user reports, and that they take appropriate measures to test the list on a reasonable sample of content already present on the service (and review any content identified by that testing to identify false positives).

<sup>272</sup> Services will also need technology in place to test the keyword list on the service, but we do not anticipate this resulting in additional significant costs beyond those incurred by a service in setting up the keyword detection technology.

to develop keyword lists or to provide broader keyword detection solutions.<sup>273</sup> Keyword management by third parties may enable services to moderate and manage custom keyword lists, using specific terms relevant to this particular offence. The cost of procuring external services may range depending on the level of support required, with one third party providing keyword management services for between £2000 to £5000 a month depending on how advanced the requirements are.<sup>274</sup>

- 14.262 Recognising the breadth of content present on in-scope services, we would expect the development of a keyword list in accordance with this option could involve a small number of weeks of work on a full-time-equivalent basis. This may involve input from regulatory staff, as well as legal and experts in ICT. Overall, for most services we expect these costs to be in the thousands, though this could be higher for the largest and most complex services. It is likely that services on which content amounting to an offence concerning articles for use in fraud is more prevalent, and services with greater volumes of regulated user-generated content, would incur greater costs in compiling and testing the keyword list respectively. However, we would expect greater benefits to come with those greater costs - in particular, by ensuring the accuracy and effectiveness of the keyword list, so far as possible.
- 14.263 This option also specifies that services review their keyword list at least every six months. The costs associated with this review will again likely vary by service.<sup>275</sup> We would expect the costs to be greater for services whose existing keyword list is (relatively speaking) less accurate and less effective, and for services on whom content amounting to the supply of, or offer to supply, articles for use in fraud is more prevalent. In both cases, we would expect greater benefits to also be associated with those greater costs. In any event, we would generally expect such costs to be materially lower than those incurred when first compiling the list.

#### *Maintaining keyword detection technology*

- 14.264 Based on our standard cost assumptions set out in Annex 14, we estimate maintaining the technology will be approximately 25% of the initial set-up costs. To ensure we are fully capturing potential costs, we assume again that this requires a similar time input from software engineers and other skilled professionals. Based on this, we estimate labour costs relating to basic ongoing maintenance of the tool to range from £7,000 to £80,000 annually. Maintaining the system involves activities such as applying updates, adjusting parameters, and ingesting new keyword lists. Again, we expect costs for larger and more complex services to be on the upper end of this range and for the largest services, they may go beyond this estimate.

#### *Quality assurance of the keyword list and technology*

---

<sup>273</sup> One online service noted that they build their keyword list through a combination of in-house efforts (with data scientists and human reviewers), alongside third-party vendors that help ingest data (for example information on compromised card details already being sold on the dark web). [CONFIDENTIAL X]. The use of key term detection was in relation to frauds such as refunds, chargeback and ticket scams, not in relation to these particular offences.

<sup>274</sup> [CONFIDENTIAL X].

<sup>275</sup> One service provider noted that if they find that a certain key term is coming up within a certain abuse vector they can easily and quickly add the key term to their keyword list. In particular, they explained that this can be done on a weekly basis. However, where there is greater complexity and the service seeks to validate the returns and test for false positives, they have to carry this testing out with data scientists on a monthly or longer basis. [CONFIDENTIAL X].

- 14.265 We are proposing that services will also have to ensure that human moderators review a reasonable sample of detected content for the purposes of assuring the quality of the keyword list and technology. These costs will vary depending on whether they are outsourced or review is done internally.
- 14.266 As a general rule, the greater the volume of illegal content on the service and/or the greater the volume of false positives found in the prior review period, the higher the potential cost associated with human review. However, we would expect the benefits to be greater in these cases too. For example, whilst a greater volume of false positives in the prior review period should result in greater human review, the information obtained by services from that review will enable them to improve the accuracy and effectiveness of the technology. We would expect this to lead to a reduction in the volume of false positives in future review periods, which should both reduce subsequent human review costs and the amount of content detected that does not amount to an offence concerning articles for use in frauds. We also note that, taking account of the nature of the offence and the fact that any detected content would be in text form, we would expect human reviewers to be able to review flagged content relatively quickly.

#### *Content moderation*

- 14.267 This option provides that, once detected by keyword detection technology, content should be considered by services in accordance with their internal content moderation policies. This means services can consider the content in a way that is appropriate and cost effective for them.
- 14.268 The cost of resourcing services' content moderation systems and processes (so as to consider content detected by the keyword detection technology) is likely to vary by service and size, depending on the details of their internal content moderation policies and the volume of detected content.
- 14.269 The greater the volume of detected content, the greater the costs are likely to be (all else being the same). Where that content does in fact amount to an offence concerning articles for use in frauds, then we would also expect the benefits to be greater, as it will ultimately result in the identification and takedown of more priority illegal content.
- 14.270 However, we recognise that greater costs arising from the consideration of content that is not in fact illegal do not necessarily result in greater benefits (and that this may in fact divert the attention of services' content moderation systems and processes from other potentially illegal content). It is for this reason that we have incorporated within the draft Code of Practice a number of measures to secure the accuracy and effectiveness of the keyword detection technology and keyword list, as far as possible, and we are not specifying precisely how services should prioritise review of detected content compared to review of other illegal content.

## **Rights impacts**

- 14.271 This section discusses the potential impacts of the keyword detection measure on individuals' rights to privacy under Article 8 ECHR and to freedom of expression under Article 10 ECHR. As explained earlier in this Chapter, these are qualified rights, interference with which may be justified on specified grounds such as the prevention of crime.
- 14.272 In considering whether impacts on these rights are proportionate, our starting point in line with the approach adopted elsewhere in this Chapter is to recognise that Parliament has

determined that the Act provides for certain offences concerning articles for use in frauds (including the supply of, or offer to supply, articles for use in frauds) to constitute priority offences, and that a substantial public interest exists in measures which aim to reduce the prevalence and dissemination online of content that amounts to those offences. That public interest relates, in particular, to the prevention of crime, the protection of health and morals, and the protection of the rights of others.

- 14.273 The detection and removal of such content acts directly to prevent crime in a number of ways, such as by deterring users from posting such illegal content or by preventing other users from accessing it (and using it in order to commit further offences, many of which are themselves also listed as priority illegal offences under the Act). It similarly acts to protect public morals, including by preventing users inadvertently encountering content online which encourages or facilitates the commission of fraud offences.
- 14.274 The removal of such content (which can include, or offer to supply, individuals' stolen personal and financial credentials) can also act directly to protect the rights of individuals in relation to their personal data and under Article 8 ECHR. We place considerable weight on these positive impacts.
- 14.275 We turn now to consider adverse impacts on users' privacy and freedom of expression.

### Users' privacy

- 14.276 This option only recommends the use of keyword detection technology (and testing of the keyword list) on content that is communicated publicly by means of the service. As such, the creation of the keyword list and subsequent implementation of the keyword detection technology should not affect the confidentiality of communications.
- 14.277 However, as noted above, we consider that there could be some circumstances in which a person may still have a reasonable expectation of privacy (for the purposes of Article 8 ECHR) in relation to content which is nonetheless considered to be communicated publicly for the purposes of the Act. In addition, the processing of text – both in order to test the keyword list on the service, and when implementing keyword detection technology in accordance with the option – may involve the processing of personal data of individuals.
- 14.278 Insofar as a service processes individuals' personal data for this purpose, any interference with users' rights to privacy under Article 8 ECHR would not be significant. Such processing will also need to be undertaken in compliance with relevant data protection legislation (including, so far as the UK GDPR applies, rules about processing by third parties or international data transfers).
- 14.279 Review of detected content by human moderators, including those employed by a contracted third party, involves more significant potential impacts on privacy. The impact on users generally would be limited, as the moderators would only access content that the user has already communicated publicly. There may be a more substantial impact on the privacy of any victims whose personal or financial information is included in the content reviewed by moderators. However, that review (and the associated interference) is for the purpose of ensuring the accuracy and effectiveness of the keyword list and keyword detection technology, and thus important to ensuring the proportionality of this option. It also helps to protect the rights of other victims or future victims for this reason, and should also help to limit any potential impacts on users' freedom of expression. Our provisional view is that a less intrusive approach (as regards victims' privacy) would not be sufficient.

## Users' freedom of expression

- 14.280 Insofar as content amounting to a priority offence regarding articles for use in frauds is correctly detected and taken down, any restrictions in relation to that content are clearly justified to protect overriding public interests.
- 14.281 Potential interference with users' freedom of expression arises insofar as content detected by services deploying keyword detection technology in accordance with this option does not amount to a priority offence regarding articles for use in frauds, but is wrongly taken down on the basis that it does. There could also be a risk of a more general 'chilling effect' if users were to avoid use of services which have implemented keyword detection technology in accordance with this option. However, we do not consider that any such effect would be significant, given that many UK users already use services which have implemented keyword detection technology and it is only recommended on content communicated publicly.
- 14.282 The design of this option includes important safeguards that mitigate against the risk that content which does not amount to a priority offence concerning articles for use in frauds is detected. These include that:
- a) services use a keyword list that contains only keywords which (either on their own or in combination with other keywords on the list) could not reasonably be expected to be used on the relevant service except in relation to the commission of an offence concerning articles for use in frauds. We are also proposing to recommend that services test their keyword list, review it at least every six months, and secure it from unauthorised access, interference or exploitation; and
  - b) human moderators review a reasonable sample of detected content, as a form of quality assurance, and that these reviews be taken into account by the service when it reviews the configuration of the technology and the keyword list.
- 14.283 However, as explained in Annex 15, we recognise that keyword detection technology deployed in accordance with this option may identify content that does not amount to such an offence. Further, the measure does provide flexibility in several respects, including as to the words used in the keyword list, as to the configuration of the keyword detection technology, and as to the steps taken by the service with detected content. There could therefore be variation in the impact on users' freedom of expression arising from services' different implementations of the technology and different approaches to moderation and take down of any detected content. Implementations that substantially impact on freedom of expression, including the automatic take down of detected content, could be in accordance with the measure in our Code of Practice.
- 14.284 Impacts on freedom of expression could in principle arise in relation to the most highly protected forms of content, such as religious or political expression. However, we consider there is unlikely to be a systematic effect on these kinds of content: for instance, such content would be unlikely to be particularly vulnerable to false positive detection, and whether or not such content were, incorrectly, subject to takedown would depend on the approach to content moderation adopted by the service, rather than the content's detection by the keyword detection technology in and of itself.
- 14.285 We recognise that keyword detection technology implemented in accordance with the option detailed here may, in principle, detect legitimate content about articles for use in frauds online (for example, news publisher or journalistic content which considers the use by offenders of keywords to facilitate its dissemination online). However, the Act includes some

protections for news publisher content which should limit the extent of any interference with such content arising from this option. Specifically, section 55(2)(g) of the Act provides that news publisher content (as defined in section 55(2)(8)) is excluded from the definition of regulated user-generated content. This option does not therefore recommend the use of keyword detection technology on news publisher content.

- 14.286 The Act itself may also provide further protection to users whose content is wrongly taken down following detection by the keyword detection technology we are recommending. Where a service takes down content on the basis that it is illegal content, complaints procedures operated pursuant to section 21(2) of the Act allowing for the relevant user to complain and for appropriate action to be taken in response may also mitigate the impact on users' rights to freedom of expression.

## Provisional conclusion

- 14.287 As set out above, the dissemination of content online which amounts to an offence concerning articles for use in frauds can cause very significant harm. It is a priority illegal offence itself, and can facilitate a number of other priority illegal fraud offences (online and offline) which often result in both financial and emotional harm to individuals affected.
- 14.288 Our analysis also suggests that the use of standard keyword detection technology would be an effective means to proactively identify such content at pace and at scale when deployed using a suitable keyword list and appropriate technical parameters. It is an established, albeit relatively basic, form of technology already used by online services in a range of contexts.
- 14.289 Our proposed measure is discussed in more detail at paragraphs A15.81-A13.125 of Annex 15 and set out in draft in Annex 7. We recognise that it might impose significant costs on some services and the potential interference that the use of the technology (in accordance with our proposed measure) could have on users' rights to freedom of expression within the law and to privacy.
- 14.290 The severity of the harm we would be addressing, and the benefits that would result from diminishing this, is sufficiently great that we provisionally consider it would be proportionate to recommend that at least some services deploy keyword detection technology to identify content likely to amount to a priority offence regarding articles for use in frauds.
- 14.291 Our provisional view is however that it would not be proportionate to apply this measure to all user-to-user services. We have therefore carefully considered to which services it would be appropriate to apply it in the Illegal Content Code of Practice, taking into account both the risk of harm presented by services of different kinds and sizes, and the size and capacity of service providers.

## The levels of risk and severity of the potential harms

- 14.292 Our draft Risk Assessment Guidance recognises that not all services will have the same risk of fraud, and sets out the factors which we consider indicate that a service should be assessed as being at low, medium or high risk for harm from fraud and financial service offences. Illegal content risk assessments therefore offer a potential means to target the measure at services with a higher risk (and where we would expect the measure to carry higher potential benefits).

- 14.293 Given the severity of the harm from content which amounts to an offence concerning articles for use in frauds, our provisional view is that our proposed measure should be focused on those services that assess themselves as having a medium or high risk for fraud.
- 14.294 We recognise that offences concerning articles for use in frauds are only some of the priority illegal offences related to fraud identified in Schedule 7 of the Act. However, we would expect (taking account of the risk factors set out in the Risk Assessment Guidance) that services that assess themselves as medium or high risk for fraud would in principle also be medium and/or high risk for content which amounts to an offence concerning articles for use in frauds. We are however particularly keen to receive stakeholder evidence on this point.
- 14.295 We do not however consider that it would be proportionate to apply this measure to all services with a medium and/or high risk for fraud, regardless of size. We are concerned that this could bring into scope a very large number of services, particularly for medium risk, and do not at this stage consider that we have sufficient evidence to demonstrate that this would be proportionate. As explained earlier, we have more limited evidence about the capacity of smaller services to implement keyword detection technology and to maintain this technology as criminals shift methodology. Whilst we recognise that the costs of implementing our proposed measure may be lower for smaller services, we recognise that those costs may be a larger proportion of the revenues of smaller services. We are also mindful of the possibility that imposing this measure on smaller services may risk creating a barrier to entry and therefore to innovation and competition.
- 14.296 We have therefore considered thresholds at which this measure applies, which would be proportionate for services with a medium and/or high risk of fraud.

### Size and capacity of the provider

- 14.297 Taking account of our evidence on costs and the likelihood of harm amounting from offences concerning articles for use in frauds, our provisional view is that it is proportionate to apply the measure outlined above to large U2U services which are medium or high risk for fraud in their risk assessment.
- 14.298 We understand criminals engaging in fraudulent activity, including offences concerning articles for use in frauds, tend to target large platforms with a wider reach as this can increase their visibility and potential revenue generation. Further, we would expect the benefits of preventing the dissemination of content which amounts to an offence concerning articles for use in frauds on large services to be particularly high given their wider reach and increased visibility. A measure which applies to large services could therefore have a significant impact on the ability of offenders to find articles for use in frauds online. Whilst we recognise that our understanding of costs is more limited in relation to this measure than others, our provisional view is that large services should be capable of bearing the costs of this.
- 14.299 We recognise that excluding smaller services from this measure could result in displacement, whereby offenders move to smaller services with a medium or high risk for fraud. For this reason, and given the severity of the harm involved, we think there is a case in principle to widen the application of this measure. At this stage, however, and recognising that keyword detection technology can be a potentially costly tool, we do not consider we have the evidence to do so.



## Summary

- 14.300 Given the severity of the harm from content which amounts to an offence concerning articles for use in frauds, our provisional view is that it is proportionate to recommend the use of standard keyword detection technology to identify such content in our Illegal Content Code of Practice in relation to large U2U services that are medium or high risk of fraud.
- 14.301 Our proposed measure has been designed with the potential impacts on users' freedom of expression or privacy in mind and incorporates some important safeguards to mitigate them, as outlined above. We recognise however that keyword detection technology implemented in ways consistent with our measure could result in some content being wrongly taken down as amounting to an offence concerning articles for use in frauds.
- 14.302 Our provisional view is that any interference with users' rights to privacy under Article 8 ECHR and to freedom of expression under Article 10 ECHR are justified by the substantial public interest in the prevention of crime, the protection of health and morals, and the protection of individuals' personal data that the measure is designed to achieve, and are proportionate to the anticipated benefits from disrupting the dissemination of illegal content regarding articles for use in frauds through the proposed measure. Our provisional view is that there is not a less intrusive way of achieving these aims. Further, as explained above, whether or not such content were, incorrectly, subject to takedown would depend on the approach to content moderation adopted by the service, rather than the content's detection by the keyword detection technology in and of itself.
- 14.303 We welcome input and evidence from stakeholders on our proposed approach, particularly on the costs associated with it.

## Illegal financial promotions and investment scams standard keyword detection

---

### Harms or risks that keyword detection seeks to address

- 14.304 Criminals openly promote fraudulent investment scams and financial promotions through social media posts, preying on the desire of consumers to invest in pensions, mortgages and investment schemes to defraud consumers and make a financial gain. These frauds use a range of methodologies including social engineering, impersonation of legitimate investment companies, fake profiles, and fake claims.
- 14.305 Financial promotion ('finprom') scams and investment scams can harm service users in numerous ways. They can result in financial harm to the user and their family, as well as emotional and psychological harm. The implications can be severe and sometimes devastating, depending on the specific circumstances. For example, if an individual has lost their life savings because of an investment scam, this can have a direct impact on the individual's emotional and psychological wellbeing, both during the event and after.
- 14.306 These scams also have an impact on legitimate businesses if they have occurred because of cloned content, impersonation or misuse of credentials. The reputational risk for legitimate businesses (including the risk, which may be particularly acute for SMEs, that they lose current and prospective customers) may have a knock-on effect on the success of the business.

14.307 Research by Ofcom has shown that social media services are being used by fraudsters to expose consumers to fraudulent financial promotions which misleadingly guarantee profits and no risk of loss.<sup>276</sup> In February, the Financial Conduct Authority (FCA) confirmed that social media was a major focus for its work combatting misleading financial promotions, and that the use of social media influencers ('fin-fluencers') to endorse and market illegal financial promotions was an area of growing concern.<sup>277</sup> In May, Natwest reported that fraudsters were using fake celebrity endorsements on social media to steal millions of pounds via content promoting investment scams.<sup>278</sup>

## Options

14.308 In light of the above, we have considered whether to include in our Code of Practice a recommendation that services deploy keyword detection technology to proactively identify content that amounts to an offence concerning illegal financial promotions and investment scams<sup>279</sup>, and to reduce the risk of services being used for the commission or facilitation of such promotions and scams.

14.309 To aid our assessment, we have considered whether a measure which is similar to the one we are proposing regarding articles for use in frauds (above) might be appropriate and proportionate. Crucially, it would rely upon services obtaining access to a list which contains only keywords which would not reasonably be expected to be used except in relation to the commission of financial promotion scams and investment scams.

## Provisional conclusion

14.310 We recognise that financial promotion fraud and investment scams are a serious harm online with many implications (e.g., financial, emotional, reputational) for those that are affected by them. It is a harm that is prevalent on a number of platforms and growing quickly.

14.311 We note that some services are already taking measures to counteract this harm. Some of those measures involve using keyword detection (sometimes in conjunction with other measures).

14.312 However, unlike the keyword detection measure regarding articles for use in frauds, we are concerned that the terminology used for both legitimate and illegal finprom is used in many other contexts. In particular, though there are words and phrases associated with financial promotion fraud, we understand that some of these are very common. We recognise that this can negatively impact the degree of accuracy and effectiveness of keyword detection technology used in this context, and have significant financial cost implications (for example,

---

<sup>276</sup> Our research revealed that 40% of respondents had experienced an investment, pension or 'get rich quick' scam, and that social media was the second most likely channel for respondents to encounter a scam or instance of fraud. Source: Ofcom, 2023. [Online Scams & Fraud Research](#). [1 August 2023]

<sup>277</sup> FCA, 2023. [Financial watchdog blocks thousands of misleading ads](#). [accessed 20 March 2023].

<sup>278</sup> Natwest's Celebrity Scam Super League table, cited Dragon's Den host Peter Jones, Sir David Attenborough, Piers Morgan, Jeff Bezos and Martin Lewis as the most commonly used images in social media investment scams. NatWest Group, 2023. [Dragon's Den star exploited by scam-ad criminals](#). [accessed 10 June 2023].

<sup>279</sup> Paragraphs 31 and 32 of Schedule 7 of the Act set out financial services offences that should be considered as priority offences. These include misleading statements and misleading impressions under sections 89 and 90 of the Financial Services Act 2012.

where services need to manage a large volume of identified content, including potentially large volumes of false positives), and significant freedom of expression implications.

14.313 Therefore, at this stage we are not proposing to recommend the use of keyword detection technology to identify investment and financial promotion scams. We do however recognise the risk and severity of this harm, and are keen to receive evidence from stakeholders on the types of measures that services are (or could) adopt to tackle illegal financial promotions online, including evidence on the accuracy, effectiveness and potential for bias of these, as well as the associated financial costs.

## CSAM standard keyword detection

---

### Harms or risks that keyword detection seeks to address

14.314 As described in the Register of Risks, child sexual abuse material is prevalent online. When this material is uploaded or shared (including by links), perpetrators will often name the file or include words in any accompanying content that enable the material to be discovered by those seeking to view CSAM. The words used can vary from commonly used words or phrases to much more specific or coded words or phrases which are known within offender communities.

### Options

14.315 We recognise that keyword lists of words or phrases associated with CSAM can be used to help detect CSAM shared or uploaded on services, and that some services are using these lists to help reduce the prevalence of this content on their services. We have therefore considered whether to include in our Code of Practice a recommendation that services deploy keyword detection technology to proactively identify CSAM.

14.316 We outline below our understanding of the way in which words or phrases can be associated with CSAM, the availability and content of existing keyword lists, the type of keyword matching that could be used, and the action taken when a match is detected:

- CSAM keyword lists are provided to services to support CSAM detection and removal by NGOs. Lists may also be provided to services by dedicated technology solutions companies. In addition, we understand from responses to Ofcom's 2022 Illegal Harms Call for Evidence that some services (particularly adult, gaming and some social media services) maintain their own lists of 'banned words' or terms that violate their terms of service.<sup>280</sup>
- The words and phrases used in CSAM keyword lists vary in terms of specificity. Some keyword lists include very specific terms, which are intended to only produce CSAM results. By contrast, we understand there are other keyword terms which are used in many different contexts, and therefore are not necessarily strong indicators of illegal CSAM content.
- There are different variations of keyword matching, and this requires consideration of the potential increased effectiveness of fuzzy matching in detecting CSAM (and thus reducing the circulation of CSAM online and the harms

---

<sup>280</sup> OnlyFans response to 2022 Illegal Harms Call for Evidence; Mumsnet response to 2022 Illegal Harms Call for Evidence.

caused) while returning a greater number of false positives, and whether these false positives can be managed by reviewers, as well as consideration of the potential reduced effectiveness of direct matching in detecting CSAM whilst detecting fewer false positives, and thus having a lower impact on privacy and freedom of expression and likely requiring a lower level of review and management.

- Services will generally deploy keyword lists to search for multiple keywords or combinations of keywords to flag user-generated content for review. We currently have limited evidence about the specifics of this in terms of the parameters needed to return a match; for example, the rules used to flag content using keyword combinations and how this is implemented depending on the particular term's correlation with CSAM.

14.317 We consider that further evidence and understanding is needed in some areas such that we are not proposing a specific measure at this time. This is due to the following factors:

- There may be variation in the specificity of, and terms included in, keyword lists. For instance, while detection of specific keywords used virtually exclusively in the context of CSAM might, in principle, be highly accurate in detecting CSAM and have little potential for impacts on users' freedom of expression arising from taking down detected content, detection of keywords that might be indicative of CSAM, but were also in common usage for benign purposes, could have significant potential impacts on freedom of expression. This would require Ofcom to consider the incorporation of safeguards in any Code measure to protect freedom of expression (such as human review of detected content).
- Evidence from industry indicates that keyword lists are used alongside human review, and although this can help to ensure that content is not incorrectly removed from a service, it is likely to have significant cost implications for services and may also have freedom of expression and privacy implications if content is incorrectly flagged as being CSAM. The main concern regarding cost would be managing the potentially material flow of false positive results. Depending on the list used and the specificity of the terms on the list, the false positive rate may be high. Therefore, significant moderator time would be diverted from other tasks to reviewing these false positives and managing complaints.

## **Provisional conclusion**

14.318 Whilst we know that forms of keyword detection for CSAM are deployed across a number of services, we consider that further evidence is needed about the quality, content (including specificity) and availability of existing keyword lists. We also have limited information on the type of keyword detection that services deploy and how this is deployed (whether to automatically block or flag content, used in conjunction with content review, or to aid machine learning). We are also conscious of the potential impacts on freedom of expression of applying this technology, given the commonly used words closely identified with this harm.

14.319 We therefore do not intend to include a measure in our Code of Practice recommending that services use keyword matching to detect CSAM at this time. However, we welcome input and evidence from stakeholders on possible approaches relating to CSAM keywords as we

continue to build our research and evidence base with a view to considering whether to make a recommendation in this area in future.

## Other policy options for standard keyword detection

---

### A broader recommendation for standard keyword detection

- 14.320 We are not proposing to recommend the use of standard keyword detection to identify other kinds of illegal content (beyond articles for use in frauds, finprom fraud, or CSAM content) at this time.
- 14.321 We note that standard keyword detection could have an important role in reducing harm in some areas, including drugs and psychoactive substances and firearms and weapons offences, and perhaps threats and harassment. However, we currently have limited information on how services can and do use keyword detection to address these harms, and the likely effectiveness and cost of applying this technology in those contexts. We would welcome more evidence on the application of keyword detection technologies to these and other offences.

## Use of AI to detect previously unidentified illegal content

---

### Introduction

- 14.322 Services use AI and Natural Language Processing (NLP)<sup>281</sup> programs to detect and remove some types of illegal and violative content, including content that has not previously been identified as illegal or violative content.
- 14.323 Some services use their own AI tools, some purchase these tools from third parties, and others use a mixture of internal and external tools.
- 14.324 These tools can be used in many different ways. The services below illustrate some of the different applications of AI/NLP:
- OpenAI is an AI research organization that offers NLP tools for content moderation, such as language models, text classification, and sentiment analysis.
  - WebPurify provides content moderation services for social networks, dating websites, and gaming platforms. It uses machine learning and NLP to analyse user-generated content.
  - Sift provides fraud detection and content moderation services for e-commerce and financial services companies. It uses machine learning and NLP to analyse user-generated content and detect fraud.
  - Many cloud service providers also offer these services to their customers including Amazon Web Services, Microsoft Azure, and Google Cloud Platform.
- 14.325 We understand that AI is being applied to combat a number of online harms, and there is, in some instances, considerable effort and investment underway to develop and implement

---

<sup>281</sup> Natural language processing (NLP) is a subfield of computer science and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process, analyse and understand large amounts of natural language data.

technology to aid in the detection and removal of illegal content. For example, we understand that AI can be used to detect previously unidentified CSAM, either which has not previously been uploaded ('first-generation' CSAM), or which has previously not been detected. We understand that this involves machine learning models to indicate the likelihood that a piece of content is or is not CSAM. AI can also be used to capture behavioural characteristics of those attempting to groom children, and is used to disrupt CSEA offending. AI and machine learning can also make use of other tools to train systems; for example, the use of keyword lists to kick-start machine learning models.<sup>282</sup>

## Provisional conclusion

- 14.326 We recognise that these technologies are increasingly used by services, and are likely to be an important tool in services' arsenal for tackling illegal content and other harmful content.
- 14.327 Nonetheless, we know relatively little about how these technologies are applied in practice, their effectiveness in tackling harm, and the likely costs involved. We further note that some applications of this technology are quite embryonic, and in some instances there are potential concerns regarding bias and interference with users' freedom of expression.
- 14.328 We are not therefore proposing to recommend the use of AI to detect previously unidentified illegal content in our first illegal content Codes of Practice. This should however in no way be interpreted as precluding or discouraging services from adopting these technologies voluntarily. As noted in the introduction to this chapter, we welcome innovation and investment in safety technologies. We intend to conduct further research and analysis on the use of AI and NLP, and may explore recommending these in future iterations of our Codes of Practice.

## Use of systems which score and track risk cumulatively

---

### Provisional conclusion

- 14.329 We are aware from our desk research<sup>283</sup> that some services use systems which take contextual signals and risk factors<sup>284</sup> into account when assessing whether to take action in relation to content or user accounts, or in deciding what action to take. Our desk research also suggests that various third parties provide the tools needed to continuously monitor for red flags as part of a cumulative approach to monitoring risk online.<sup>285</sup>
- 14.330 We describe these as 'cumulative risk scoring systems', as content or accounts are given a risk score that reflects multiple factors such as internal user reports and / or third-party intelligence. The risk score can be used to flag high risk content or accounts for review or for automated moderation decisions.

---

<sup>282</sup> EA response to Ofcom's 2022 Illegal Harms Call for Evidence; Mumsnet response to 2022 Illegal Harms Call for Evidence.

<sup>283</sup> For example, TikTok (Han, E.), 2021. [Advancing our approach to user safety](#). [accessed 21 March 2023].

<sup>284</sup> By 'contextual signals' and 'risk factors' we are referring to specific online behaviours that suggest there is suspicious or atypical behaviour occurring. For example, this may include the use of terminology that is known to be associated with fraud (i.e., the sale of stolen identity credentials, suspicious IP activity, accounts frequently flagged by other users etc).

<sup>285</sup> Fraud.net, 2023. Account AI: Comprehensive Risk Analysis Across the Entire Lifecycle of Accounts. [accessed 23 June 2023]; and Microsoft, 2023. [What are risk detections?](#) [accessed 23 June 2023].

- 14.331 In principle, such systems could enable services to comply with their safety duties about illegal content more effectively by ensuring that any actions take account of a range of relevant contextual information. Evidence from the banking and cyber security sectors suggests that a cumulative risk scoring system can enable businesses to make best use of the data available to them internally and externally to detect suspicious and malicious activity.<sup>286</sup>
- 14.332 We consider that cumulative risk scoring systems could provide various benefits for tackling illegal harms such as fraud, drugs and weapons offences, child sexual exploitation and abuse, terrorism, and unlawful immigration. We recognise however that there is significant complexity involved in these systems, and that there could be adverse impacts on user privacy or freedom of expression if the operation of the system were to result in inappropriate action being taken against content or user accounts. We have limited evidence on this at present. As a result, we are not proposing to include a recommendation that services use cumulative risk scoring systems in our Codes of Practice at this time.
- 14.333 We intend to engage with services in due course to gain a better understanding of which services already use measures of this nature to tackle illegal content and to develop our understanding of the associated risks, with a view to considering whether to recommend the use of cumulative risk scoring systems in future iterations of our Codes of Practice.

---

<sup>286</sup> Thales, 2023. [Fraud detection in banking with IdCloud risk management](#). [accessed 20 May 2023].



# 15. Automated Search Moderation

## What is this chapter about?

In our Search Moderation (Search) chapter, we explained our proposals in relation to the measures services should take to set up their search moderation systems in a manner consistent with the safety duties. Search services may use automated tools to make moderation processes more effective at identifying and taking action in relation to illegal and violative content. As these tools enable services to moderate large numbers of search results at pace, they can be critical to services' attempts to reduce harm. This chapter focuses in detail on automated moderation tools and assesses what automated tools our Codes should recommend search services use.

## What are we proposing?

We are making the following proposal for all general search services:

- **Ensure that URLs which have been identified as hosting CSAM or as being part of a website entirely or predominantly dedicated to CSAM are deindexed from the search index of a relevant service.** Services should source an appropriate list of CSAM URLs from third parties with expertise in the identification of CSAM and which meet other identified criteria. The list should be regularly monitored to identify new CSAM URLs and take steps to deindex, and reinstate CSAM URLs that have been removed from the list.

## Why are we proposing this?

The circulation of CSAM online is increasing rapidly. The evidence presented in volume 2 shows that perpetrators use search services to access CSAM and the NCA has shown that it is possible to find CSAM within three clicks on some major search services. As we explained above, child sexual abuse and the circulation of CSAM online causes significant and potentially 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.

## What input do we want from stakeholders?

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

## Introduction

---

15.1 In Chapter 13 we discussed search moderation and set out the measures we propose to include in our illegal content Codes of Practice in relation to search services' search moderation functions and policies. We explained that, in drafting these measures, we proposed not to be prescriptive as to services' design of their search moderation systems and processes, but instead to set out factors to which they should have regard. Chapter 13 defined a search service as set out in the Online Safety Act and set out our proposed definitions of types of search service to enable targeting of measures in paragraph 13.2. The types are: general search services – which includes large general search services, small

general search services and downstream general search services – and vertical search services.

- 15.2 Separately from that general approach, in this chapter we consider whether to include any measures recommending the use of automated systems and processes (including technologies) by search services to reduce the risk of users encountering search results that link to illegal content. Such systems and processes can play a crucial role in supporting search services' compliance with their illegal content safety duties under section 27(2) and (3) of the Act, given the enormous volumes of content on websites or databases that may be searched by search engines.
- 15.3 We consider the use of automated content moderation (ACM) for user-to-user (U2U) services in Chapter 14. As explained in that chapter, our proposals in relation to U2U services relate to the use of well-established automated tools which are already used by many services to reduce harm relating to illegal content. Search services are distinct from U2U services in that they do not facilitate the sharing or uploading of content by the user of the service, but rather facilitate access to more than one website or database and thereby can act as a gateway to illegal content that is present elsewhere online. Because of this difference, we give separate consideration to how automated moderation tools could apply to search services.
- 15.4 Potential tools include automated (or part-automated) systems for **deindexing** (or delisting), **downranking**, or other forms of prioritisation for search results, as described in Chapter 13. **Deindexing** is one of the primary means by which general search services can control the visibility of what appears in search results and, as a result, minimise the risk of users encountering content.
- 15.5 **Deindexing** tools can automate the review of listings appearing in a search by comparing material contained in a search index against a database of known illegal content. Material in the index that matches existing content in the database can then be flagged for further review or automatically deindexed.
- 15.6 A search service can also downrank search results by altering the ranking algorithm to ensure that a particular piece of content appears lower in the search results and is therefore less discoverable to users, thereby minimising the risk of users encountering illegal content via search results. As with deindexing, this process can be automated.
- 15.7 Depending on what harm these automated tools are applied to, and in what way, their accuracy, effectiveness and degree of bias can vary. They can therefore have a significant impact on user rights, in particular freedom of expression and privacy. They can also incur significant costs, varying depending on the nature and complexity of the technology and how it is applied. Due to this, we have assessed the tool we consider in detail.
- 15.8 Technologies which analyse search content to assess whether it is illegal content (or other content of a particular kind) are "content identification technologies" (as defined in section 231 of the Act) and will (subject to an exception, not relevant here) fall within the definition of what the Act refers to as "proactive technology". Paragraph 13 of Schedule 4 to the Act contains a number of constraints on Ofcom's ability to recommend the use of "proactive technology" in codes of practice. These include that:

- Ofcom may not recommend in a code of practice the use of the technology to analyse user-generated content communicated privately, or metadata relating to user-generated content communicated privately.<sup>287</sup>
  - When deciding whether to include a proactive technology measure in a code of practice, we must have regard to the degree of accuracy, effectiveness and lack of bias achieved by the technology.
  - A proactive technology measure may be applied to services of a particular kind or size only if we are satisfied that the use of the technology in question by such services would be proportionate to the risk of harm that the measure is designed to safeguard against (taking into account, in particular, the risk profile relating to such services).
- 15.9 Where we discuss (but do not propose to recommend) measures which may be effective in reducing risks of harm, this is principally because of our limited evidence base at this stage and should not be read as an indication that we consider taking such measures to be disproportionate. We welcome innovation and investment in safety technologies, and plan to consider further automated search moderation measures for future versions of our Codes.
- 15.10 More broadly, there is a scarcity of evidence on how priority illegal content or other illegal content can be accessed via search services. Therefore, the following recommendations are largely reflective of current industry standard practice. We plan to expand on the provisions relating to search in Codes as we build our knowledge and understanding.

## Child Sexual Abuse Material (CSAM) URL deindexing (search)

---

### Introduction

- 15.11 In principle, the risk of users encountering illegal content by means of a search service exists in relation to any illegal content available to be indexed, including all priority illegal content. For example, we understand there is evidence that search services can be used to access terrorism content.<sup>288</sup>
- 15.12 However, the evidence indicates that this risk is heightened for CSAM, by which we refer to indecent or prohibited images of children, or other material which contains advice about grooming or sexually abusing a child 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 CSAM, or which advertises the distribution or showing of CSAM. References to CSAM in this chapter are not limited to images and videos, and include written and audio content.<sup>289</sup>
- 15.13 In this chapter, we consider recommending the use of automated systems and processes by general search services to detect URLs at which CSAM is present which might otherwise appear in search results, and deindex or downrank those results.

---

<sup>287</sup> See Annex 9 for an explanation of our proposed approach to deciding whether content is communicated 'publicly' or 'privately'.

<sup>288</sup> For further information, see Volume 2: Chapter 6 - Part 2 Register of Risks (Search).

<sup>289</sup> For further detail, see Annex 10, Chapter 5. Child sexual abuse and exploitation (CSEA): Offences relating to child sexual abuse material (CSAM) of the draft Illegal Content Judgements Guidance.

## Harms this measure seeks to address

- 15.14 Research carried out by the National Crime Agency on the accessibility of CSAM via mainstream search engines found that this material could be found within three clicks.<sup>290</sup> Indeed, search engines are one of the most common means used by individuals to find CSAM.<sup>291</sup>
- 15.15 CSAM causes serious harm, whether accessed via search services or in other ways. Studies indicate a connection between viewing CSAM and going on to contact children for the purposes of sexual abuse. One study found that 37% of perpetrators who had viewed CSAM online went on to seek sexual contact with a child afterwards; 5% of perpetrators said that this was on a weekly basis.<sup>292</sup> This indicates that reducing access to CSAM may also result in a reduction of other types of child sexual abuse, such as grooming and contact abuse.
- 15.16 The Register of Risks sets out the profoundly negative impact that being sexually abused as a child has on victims and survivors. In particular, analysis by the Independent Inquiry into Child Sexual Abuse found that 88% of victims and survivors reported a negative impact on their mental health.<sup>293</sup> Child sexual abuse often also has a severe impact on physical health, including as a result of physical injury, sexually transmitted infections and pregnancy. Further, many victims and survivors report an impact on their education, ability to work and career prospects, relationships, parenting and faith.<sup>294</sup> Reducing access to CSAM online would help to reduce the number of children experiencing this severe and often lifelong harm, thereby delivering significant benefits.

## Options

- 15.17 The risk of harm posed by accessing CSAM links through search results can be mitigated by using Uniform Resource Locator (URL) detection, which is a process by which a URL (i.e., individual webpage addresses) in a search service index is matched against a URL known to host illegal content. Once a match is established, the URL where the relevant illegal content is hosted may be automatically downranked, deindexed or sent for human review before actioning. The process relies on there being a URL list.
- 15.18 In this context, a URL list is a list of URLs at which illegal content is present. Some services compile their own URL lists, but often such lists are maintained by third parties such as NGOs, law enforcement bodies, or providers of safety tools.
- 15.19 Deindexing or downranking of URLs identified as containing CSAM, such as those included in lists maintained by reputable sources like the Internet Watch Foundation (IWF), provides a means of reducing the discoverability of this content online, given the gatekeeping role of

---

<sup>290</sup> UK Government, 2020. [Interim code of practice on online child sexual exploitation and abuse](#). [accessed 26 May 2023].

<sup>291</sup> Steel, C.M.S., 2015. [Web-based child pornography: The global impact of deterrence efforts and its consumption on mobile platforms](#), *Child Abuse & Neglect*, 4. [accessed 12 June 2023].

<sup>292</sup> Insoll, T., Katariina Ovaska, A., Nurmi, J, Aaltonen, M. and Vaaranen-Valkonen, Nina., 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 12 June 2023].

<sup>293</sup> See Volume 2: Chapter 6C CSEA (grooming and CSAM).

<sup>294</sup> Independent Inquiry into Child Sexual Abuse, 2022. [The Report of the Independent Inquiry into Child Sexual Abuse](#). [accessed 28 September 2023].

search services and the extent of their use by users as a means of accessing content on the web.

- 15.20 Downranking of URLs does not prevent a user from encountering that content, as the content could still be accessed via the search service, just with greater effort and time. This could be an effective mitigation in some instances given that research suggests that users often do not go beyond the first page of search results.<sup>295</sup> However, we consider that it would not be a sufficient measure for addressing the harm caused by content that has been identified as CSAM, given the seriousness of such content. We have therefore not considered this option further in our assessment below.
- 15.21 We have instead considered a potential option that all general search services should detect and deindex known CSAM URLs.
- 15.22 Where we refer to CSAM URLs, this would include not only URLs at which indecent or prohibited images or other material is present, but also URLs which redirect or include links or otherwise direct users to such URLs (or which advertise the distribution or showing of indecent or prohibited images). It would also include URLs hosted by a domain which is entirely or predominantly dedicated to CSAM.<sup>296</sup>
- 15.23 Below we discuss how any measure related to CSAM URL detection would need to be designed. Once we have done this, we then go on to assess the impact of such a measure.

## Discussion of design of the measure

### Technology used for deindexing

- 15.24 Under the option we are considering for search services, URL detection technology would employ direct matching, which would only detect a match if a URL in a search service's index exactly matches that on a URL list of known illegal content. As this requires an exact match with previously identified URLs hosting CSAM content, the risk of surfacing links to content that is not CSAM is low, though this is dependent on the quality and accuracy of the underlying list.

### URL lists

- 15.25 We understand that many search services rely on URL lists maintained by the IWF and other child protection organisations or law enforcement bodies globally to support their CSAM deindexing efforts. We understand that these lists vary in terms of whether they include URLs to specific pages or domain-level URLs.
- 15.26 We also understand that some lists may only be available to certain organisations, such as law enforcement bodies or internet service providers (ISPs). This includes, for example, Interpol's 'Worst of' list (IWOL), which is a list of domains provided to ISPs to block access at the network level.<sup>4</sup>

---

<sup>295</sup> See for example: Höchstötter, N. and Lewandowski, D., 2009. [What users see – structures in search engine results pages](#), *Information Sciences*, 179(12). [accessed 28 September 2023]; Jansen, B.J. and Spink, A., 2006 [How are we searching the World Wide Web? A comparison of nine search engine transaction logs](#), *Information Processing Management*, 42(1) [accessed 28 September 2023]; Pan, B. *et al.*, 2007. [In google we trust: Users' decisions on rank, position, and relevance'](#), *Journal of Computer-Mediated Communication*, 12(3). [accessed 28 September 2023].

<sup>296</sup> See further paragraphs A15.73 to 77 of Annex 15.

- 15.27 In its response to our call for evidence, IWF stated that it provides “a range of technical services including a webpage blocking list (URL list)”.<sup>297</sup> Google indicated in its response to our illegal harms call for evidence that it uses URL lists maintained by both the IWF and NCMEC.<sup>298</sup> We understand that other services such as Microsoft and Mojeek also work with the IWF to deindex URLs identified to contain CSAM and pay membership fees to access these lists.<sup>299</sup>
- 15.28 We understand that some larger search services may, or would have the resource to, maintain their own list of URLs. However, due to the legal and practical risks of doing so outlined in paragraph A15.68 of Annex 15 and the current practice around deindexing of CSAM URLs, we consider it appropriate that services use lists prepared by reputable third parties.
- 15.29 We recognise that the effectiveness of the measure would depend on third party lists being developed by a person with expertise in the identification of CSAM, and on the arrangements in place to ensure the integrity of that list. Our provisional view is therefore that our recommendation should include the same criteria outlined in paragraph A15.70 of Annex 15 relating to the standards of assessment of suspected CSAM content, maintenance of the list to ensure that new CSAM URLs are included and those which no longer contain CSAM are removed.
- 15.30 This should guard against the risk of services deindexing URLs based on lists that are inaccurate or limited in scope which would make the measure significantly less effective and could have rights implications.
- 15.31 In line with the current practice of organisations like the IWF and paragraph A15.76 of Annex 15, we consider it appropriate that deindexing applies not only to URLs at which CSAM is confirmed to be present, but also URLs that redirect to such URLs and URLs hosted by domains where that domain exists predominantly or entirely for the distribution of CSAM.

## Human review

- 15.32 The appropriate level of human oversight and review depends on the accuracy of the underlying technology and URL database.
- 15.33 We understand that the false positive rate for direct matching itself is low to none. Our provisional view is that any recommendation should relate to direct matches to, and removal of, URLs contained on a list of URLs known and verified as CSAM (including directing to CSAM).
- 15.34 We would expect services to ensure that the URL lists they procure are accurately maintained and updated, and that they use the most recent version of the list made available to them.
- 15.35 We therefore would not anticipate it being necessary or proportionate for search services to use human review to check positive matches to the URL list before the URLs are deindexed from the service.

---

<sup>297</sup> Internet Watch Foundation response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

<sup>298</sup> Google response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

<sup>299</sup> Internet Watch Foundation. [Our members](#). [accessed 11 July 2023].

## Outline measure

- 15.36 We have therefore considered whether to include in our CSEA Code of Practice a recommendation that general search services use technology to deindex CSAM URLs from the service’s search index. This would involve the following elements:
- a) The use of direct matching URL detection technology;
  - b) The use of an appropriate list of CSAM URLs sourced from an organisation or person with expertise in the identification of CSAM, with arrangements in place to ensure the accuracy of the list (including when adding URLs to the list and by reviewing the list to remove URLs which are no longer CSAM URLs). The list (and any copy held for the purpose of the measure) should also be secured against unauthorised access, interference or security compromises through attacks by bad actors.
  - c) Regularly monitoring that list to identify new CSAM URLs and reverse deindexing for any URL that is subsequently removed from the list.
- 15.37 As outlined in Chapter 14 in relation to U2U URL detection, it would be appropriate in most instances for the URLs included on the list to be URLs of the specific webpage at which CSAM is hosted (rather than listing whole domains) to avoid “over-blocking” of legitimate content. In addition, however, we also consider that it would be appropriate to list at domain level where the domain is entirely or predominantly dedicated to CSAM.<sup>300</sup> This is likely to be more effective and efficient than listing each individual URL containing CSAM, given that these may alter frequently. However, we would expect services to ensure that the provider of the URL list has arrangements in place to ensure that listing at domain level only occurs in such cases.

## Accuracy, effectiveness and lack of bias

- 15.38 We consider deindexing to be an effective defence against accessing illegal and harmful content via search services. Once a URL is deindexed, it is removed from the search index and can no longer be accessed via the search service and will no longer appear in search results. At the level of the individual instance of CSAM, deindexing would ensure that users are no longer able to encounter a specific URL via the search service. More broadly, it would reduce the overall volume of CSAM that appears in users’ results, thus providing a significant hurdle to users attempting to access such material and reducing the likelihood that users will encounter such content.
- 15.39 In its response to our call for evidence, the Australian eSafety Commissioner suggested that deindexing search results could be an effective method to address harmful content.<sup>301</sup> This is supported by evidence of current industry practice among general search services in relation to CSAM specifically:
- a) In its response to our call for evidence, Google stated that “what we can do is either “delist” or “demote” content.”<sup>302</sup> Google’s transparency report shows in that in the first half of 2021 there were 596,710 URLs, and in the second half 580,380 URLs, that were deindexed from Google Search for violating policies in relation to CSAM and almost

---

<sup>300</sup> See further paragraphs A15.73 to 77 of Annex 15.

<sup>301</sup> eSafety Commissioner Australia response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

<sup>302</sup> Google response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.



500,000 reports were made to NCMEC containing 6.7 million pieces of content including images, videos, URL links and/or text soliciting CSAM.<sup>303</sup>

- b) Microsoft's 2022 Transparency report shows that from January to June 2022, Microsoft deindexed 176,125 pieces of confirmed CSAM content on Bing.<sup>304</sup>
- 15.40 A study on the effectiveness of deterrence efforts, which included deindexing, by Google and Bing compared to Yandex showed that the efforts taken by Google and Bing resulted in a 67% reduction in CSAM related queries between 2013 and 2014 in the United States, compared to Yandex which undertook no such efforts and saw no commensurate decrease.<sup>305</sup> Whilst deindexing cannot remove or eliminate the risk of encountering CSAM content via search services altogether, deindexing known CSAM contributes to the overall minimisation of the risk that users encounter CSAM by means of a search service.
- 15.41 In respect of effectiveness, a direct matching recommendation may result in 'false negatives', where a URL containing CSAM is not detected. However, we consider it to be highly effective in detecting direct matches to CSAM URLs already on a URL list.
- 15.42 Effectiveness also depends on the completeness, accuracy, and regular deployment of the URL list being used. The option we outline above includes a number of elements designed to ensure the URL list supports the effectiveness of the measure, such as ensuring that there are arrangements to identify suspected CSAM URLs and to regularly update the list, and that services compare their search index to the latest version available of the list.
- 15.43 In respect of accuracy, deploying a form of URL matching technology to detect direct matches with URLs on a list should be highly accurate. Whether a matched URL in the search index is a CSAM URL will depend on the accuracy of the URL list. The option we outline above has elements designed to ensure that CSAM URLs are accurately included in the list, as set out above. We consider these would substantially mitigate the risk of content being incorrectly identified as a CSAM URL.
- 15.44 The measure would affect all users of general search services, so there is no risk of bias in terms of the effect on users. There is a risk of bias in relation to the compilation of the URL list (not to the technology used to match and deindex those URLs). For example, addition of URLs to the list depends on where online it is, how it is detected (e.g. through AI machine learning models, web crawling, or human analysts), and the assessment of content as CSAM (e.g., age determination). This may create biases that underrepresent the scale and nature of the problem of CSAM for different ages and minority groups, but these are mitigated by the elements which promote the accuracy and effectiveness of the list.
- 15.45 We recognise that the URL list itself may be vulnerable to security compromises through attacks by bad actors. Perpetrators may attempt to attack services to make measures less effective. For example, the simpler the implementation of the technology, the higher the risk that the service can be attacked by bad actors to gain access to the URLs in question. As such, the measure includes that URL lists should be secured from security compromises, and that appropriate measures should be taken to secure any copy of the list held by or for the service.

---

<sup>303</sup> Google. [Google's efforts to combat online child sexual abuse material](#). [accessed 11 July 2023].

<sup>304</sup> Microsoft. [Digital Safety Content Report](#). [accessed 11 July 2023].

<sup>305</sup> Steel, C.M.S., 2015.

15.46 Overall, we consider that the use of this proactive technology would be effective in reducing the access to CSAM. Deployment of URL lists would therefore deliver very significant benefits in mitigating against the harms caused by the availability of CSAM online. This includes reducing the harm caused by sharing of CSAM to victims and survivors, as well as reducing intentional viewing of, or unintentional exposure to, this content and subsequent contact sexual abuse.

## Costs and risks

- 15.47 The main costs that we would expect a service to incur in applying this measure are: the cost of obtaining an appropriate URL list; and the cost of ensuring its system acts on this list to remove relevant URLs from its index and ensure that URLs that no longer feature on the list are reinstated to its index.
- 15.48 The measure would require search services to source a URL list from a third party and it is likely that a cost would be associated with this (for instance, a fee to support an NGO's work maintaining the list). For example, the IWF has an annual fee based on industry sector and company size, which can range from over £1,000 to over £80,000 per year.<sup>306</sup>
- 15.49 Services would also need to integrate the list of URLs sourced from a third party into their existing systems and regularly test their index against the latest version of the list to remove URLs to ensure that they do not appear in search results. This approach is similar to what services have already undertaken to remove content that infringes on copyright.<sup>307 308</sup>
- 15.50 The system costs are likely to include 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.
- 15.51 We estimate that implementing this type of functionality would take approximately 2 – 16 months of software engineering time, with an equal amount of non-software engineering time. Based on labour cost assumptions set out in Annex 14, we expect the initial implementation cost would be somewhere between £20,000 and £280,000, depending on the complexity and size of the existing search service system infrastructure.
- 15.52 In addition to the implementation costs, services would incur ongoing costs. Ongoing costs would include: ongoing access management to the search engine infrastructure, additional recurring software licensing costs, costs of installing new infrastructure and any recurring annual fee to the third party to access their URL list.
- 15.53 Assuming ongoing costs are 25% of the original implementation costs,<sup>309</sup> we expect the annual running costs would be approximately £5,000-£70,000 per annum. This is in addition to any annual fee to the third party providing the URL list as mentioned above.<sup>310</sup>

---

<sup>306</sup> Internet Watch Foundation. [Membership Fees](#). [accessed 25 May 2023].

<sup>307</sup> Intellectual Property Office, 2017. [Search Engines and creative industries Sign anti-piracy agreement](#). [accessed 28 September 2023].

<sup>308</sup> UK Government, 2017. [Voluntary Code of Practice on Search and Copyright'](#). [accessed 28 September 2023].

<sup>309</sup> See Paragraph A14.5 of Annex 14 for an explanation of this 25% assumption.

<sup>310</sup> We note that this fee would normally be for membership of an organisation like the IWF which would allow access to keyword list and hash matching databases in addition to the IWF. This would reduce the cost associated with this specific measure as the overall costs would be shared across different measures.

- 15.54 The costs of implementing and running this measure may vary by service due to a range of factors. We expect that services with larger indexes would take longer and may require more resources to identify and remove relevant URLs.
- 15.55 We expect most services would use almost entirely automated systems to implement this measure given the very frequent required changes to the URL list.<sup>311</sup> Services may incur additional costs where there is some form of human involvement in the process – for example, where an employee manually replaces the URL list with a new version sourced from the relevant provider.
- 15.56 Lastly, system infrastructure varies considerably from service to service, therefore we would expect the cost of implementing this measure to vary accordingly.
- 15.57 We understand that large search services (Google and Bing) already work with established organisations such as the IWF to source URL lists and remove CSAM from their indexes, therefore we would not expect these services to incur additional upfront costs to apply this measure.
- 15.58 The extent to which smaller services would incur costs applying this measure is likely to be dependent on: whether they are downstream search services and thus have limited control over their index; and, if so, whether the upstream service that supplies them has already implemented the measure.
- 15.59 Several small downstream search services such as DuckDuckGo and Ecosia buy most/all of their search results from Bing or Google and have limited control over their search index. We understand that Bing and Google already apply this measure and there are likely to be no or negligible costs to the downstream search service to ensure this measure is met.
- 15.60 However, smaller general search services that carry their own indexing may incur the additional costs of implementing this measure. There are limited numbers of these services, and we understand that the one UK-based small general search service of this type that we are aware of (Mojeek), already applies this measure. There are also small general search services based overseas (e.g., Yandex, Baidu) that may fall in scope of this measure.<sup>312</sup> At present, we do not have information on the extent to which these services already apply this type of measure and so they may need to incur the costs set out above in order to implement the measure<sup>313</sup>.
- 15.61 Finally, there may be some smaller downstream services that make use of upstream indexes that do not apply our measures.<sup>314</sup> An example might be a search service that translates a

---

<sup>311</sup> To the extent that human involvement exists, we would expect this to be limited to manually replacing the URL list with a new version sourced from the relevant provider. The actual removal of URL's would be an automated process.

<sup>312</sup> The application of this measure would depend on whether they fall in scope of the OS regime. In practice, although these services may be small in the context of the UK, they may have significant user numbers outside the UK.

<sup>313</sup> A study on the effectiveness of deterrence efforts, published in 2015, which included deindexing, by Google and Bing compared to Yandex showed that the efforts taken by Google and Bing resulted in a 67% reduction in CSAM related queries between 2013 and 2014 in the United States, compared to Yandex which undertook no such efforts and saw no commensurate decrease. This reflected the practice of Yandex at the time of publication. We do not have information on current practice. Source: Steel, C.M.S., 2015. [accessed 10 September 2023].

<sup>314</sup> For example, if they are not in scope of the online safety regime.

larger overseas search engine.<sup>315</sup> If the upstream index does not apply the measure, then under this recommendation the downstream service would be required to ensure the measure is implemented, either by making an agreement with the upstream index or adapting the index itself. Adapting the index itself could be a material cost for these type of services and could, in some circumstances, result in the withdrawal of a service.

- 15.62 There is also a small risk to the owners of any URLs that are mistakenly identified as CSAM URLs or domains. Deindexing could have very significant commercial impacts on them. However, we understand the organisations responsible for compiling the CSAM URL list have an appeals procedure that allows web page owners to request that their URLs be removed from the CSAM list. If the URL is found to not contain CSAM content, the page will be removed from the list.<sup>316</sup>
- 15.63 The extent to which this measure would erect additional barriers to new entrants would ultimately depend on the service's approach to indexing: if the service intended to import its results from an upstream provider such as Bing we would not expect it to incur any material additional costs, as we understand Bing is already in compliance with this measure. If the service planned to build its own index independently, then it would incur the costs associated with this measure and this may represent a significant barrier to entry. However, we are aware of some smaller platforms (such as Mojeek) developing their own index and successfully entering the market, which suggests that this barrier would not necessarily prevent entry for smaller platforms.
- 15.64 To our knowledge most search services make efforts to deindex CSAM material and often work with organisations such as the IWF to obtain a URL list. On this basis, we would not expect this measure to cause widespread additional costs unless services had planned to remove the functionality in order to reduce running costs. However, given the egregious nature of CSAM content it is unlikely that any services are likely to be planning to remove this functionality, and services that import their index from larger providers would be incapable of removing this functionality in any case.
- 15.65 Any search service operating in Australia that is subject to the eSafety Search Code would be required to "delist search results that surface known CSAM". To meet this requirement, relevant search services would probably anyway need to take actions consistent with meeting the measure they have proposed.<sup>317</sup>

## Rights impacts

- 15.66 As set out in Chapter 12, content moderation is an area in which the steps taken by services as a consequence of the Act may have a significant impact on the rights of individuals and entities - in particular, to freedom of expression under Article 10 ECHR and to privacy under Article 8 of the European Convention on Human Rights ('ECHR').

## Freedom of expression

- 15.67 An interference with the right to freedom of expression must be prescribed by law and necessary in a democratic society in pursuit of a legitimate interest. In order to be

---

<sup>315</sup> For example, see [Baidu in English](#). [accessed 10 September 2023].

<sup>316</sup> Internet Watch Foundation, [IWF Content Assessment Appeal Process](#). [accessed 21 September 2023].

<sup>317</sup> eSafety, [Internet Search Engine Services Online Safety Code \(Class 1A and Class 1B Material\)](#), paragraphs 7(2)(a). [accessed 21 September 2023].

‘necessary’, the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.

- 15.68 We acknowledge that the deindexing of URLs by search services potentially constitutes a significant interference with the rights of website providers to impart information and users to receive it. It also interferes in a more narrow way with the right of search services to impart information. The removal of a URL from a search service index in practice means that users will no longer be able to encounter it via that service, affecting any legal content, as well as illegal content, hosted on the URL. This interference can only be justified in circumstances where there is sufficient certainty as to the illegal nature of the content.
- 15.69 However, CSAM is an extremely harmful kind of illegal content. So far as CSAM URLs are correctly detected and deindexed, the content contained at, or linked to via, that URL either does not engage Article 10 ECHR at all or otherwise restrictions in relation to that content are clearly justified insofar as the deindexing contributes to the prevention of crime, the protection of morals, and the protection of the rights of others (in particular, the children concerned).
- 15.70 We consider that there should be few cases where links are incorrectly taken down. The option outlined above embeds a number of features to secure accuracy and therefore safeguard freedom of expression. It recommends direct matching to a list of URLs sourced from a third party with expertise in the identification for CSAM and makes further provision to ensure the accuracy of the list and its use, as follows:
- a) the need for arrangements to secure that CSAM URLs are correctly identified before being added to the list, and to review CSAM URLs on the list and remove them where appropriate (for example because the CSAM present at the URL has been taken down);
  - b) the need for the list to be regularly updated and for the service to use the latest available version; and
  - c) for both the list and any copy of the list held for the purposes of the service to be secured from unauthorised interference (which would safeguard the list from the risk of bad actors adding URLs to the list for malicious purposes).
- 15.71 We acknowledge that there could be cases where an URL has been incorrectly included on the URL list as a CSAM URL, or where a URL continues to be deindexed for a period after CSAM has been removed from it. We also recognise that there may be some interference with freedom of expression insofar as the content present at a URL includes legitimate content as well as CSAM. The option we have outlined provides for URLs to 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, but consider that this is justified to protect public interests (given the risk that users accessing such URLs will go on to encounter CSAM).
- 15.72 We considered whether it would be necessary, in order to safeguard freedom of expression, to notify the website operator that the URL has been deindexed. We are aware that The Open Rights Group has emphasised the importance of ensuring website operators are informed of down-ranking or deindexing of their URL.<sup>318</sup> While we do not dispute that being deindexed can have very significant impacts on a website operator, both in terms of

---

<sup>318</sup> Open Rights Group, [Open Rights Group response](#) to 2022 Ofcom online safety regulation consultation, page 8.

commercial impact and freedom of expression, website operators have alternative means of determining whether their URL is indexed or not, such as via Google Search Console.<sup>319</sup> We also understand that services such as the IWF do not independently notify website operators, but rather work in partnership with local law enforcement to avoid prematurely notifying website operators which could prejudice investigations and result in the removal of vital evidence.<sup>320</sup> In the circumstances, and, particularly given the egregious nature of the harm the measure seeks to address, we do not consider that we should include in our recommendation, a recommendation that URL owners should be notified.

- 15.73 However, the Act contains a duty on search services to take appropriate action in relation to complaints from an interested person whose URL has been deindexed because of a judgement that content is illegal content.<sup>321</sup> We set out in Chapter 16 our proposals for recommendations in this area. We consider that complaints procedures operated pursuant to the Act allowing for the interested person to complain and for appropriate action to be taken in response may also mitigate the impact on their rights to freedom of expression.

## Privacy

- 15.74 Our provisional assessment is that any impact on the right to privacy under Article 8 ECHR associated with this measure would be limited.
- 15.75 The process of deindexing CSAM URLs involves the identification of exact matches between URLs on a third party list and those included in a search service's index. It does not require the search service to view or analyse the content contained at or via those URLs at any stage of the automated process. Rather, the measure relies on the existing assessment conducted by a relevant third party. While we recognise that offending content contained at, or accessible via, a CSAM URL may contain personal data from which a victim or others might be identifiable, which the provider of the third party list would need to know about to know that the URL should be included on the list, we would expect those organisations to have very robust security and GDPR compliance measures in place which would serve to safeguard against any privacy risks. Victims rights overall would be safeguarded by the measure, because it would prevent their images being seen more widely.
- 15.76 The inclusion in the measure of the requirement that both the third party list provider and search service secure the URL list from unauthorised access would provide a further safeguard to the privacy rights of any individuals that might be identifiable in the content contained at the URLs.
- 15.77 As identified in the search moderation measures proposed in Chapter 13, section 66 of the Act makes provision which (when brought into force) will require providers of regulated search services to report detected and unreported CSEA content present on websites or databases to the Designated Reporting Body housed in the NCA. Unlike moderation processes which may result in the detection of CSEA content present at URLs indexed by a service, the deindexing of CSAM URLs involves only the detection of URLs and would therefore not be likely to trigger this duty.<sup>322</sup>

---

<sup>319</sup> Google, [Google Search Console](#). [accessed 12 July 2023].

<sup>320</sup> Internet Watch Foundation, [Takedown Notices](#). [accessed 1 July 2023].

<sup>321</sup> Section 32(4)(c).

<sup>322</sup> We understand that many of the third party URL list providers that would meet the criteria would likely either be a law enforcement body itself, or work directly with law enforcement (be that by receiving URLs from, or reporting URLs to, law enforcement).



## Provisional conclusion

- 15.78 Our provisional view is that search services using URL detection technology to detect matches with known CSAM URLs can be an effective means of moderating search results to enable services to proactively identify and tackle the dissemination of CSAM. As set out at paragraphs A15.63 – A15.77 of Annex 15, our provisional view is contingent on services deploying an appropriately maintained list of CSAM URLs and having sufficient oversight to ensure the process is working as intended.
- 15.79 We have considered to which services it would be appropriate to apply the measure outlined above (and which is set out in draft in Annex 8) in the CSEA Code of Practice.
- 15.80 Overall, we consider that introducing a measure that requires general search services to deindex known CSAM URLs would be both effective and proportionate for all general search services. We recognise that there could be some costs to services in applying such a measure, but we consider these costs are likely to be proportionate because:
- a) The egregious nature of the potential harm combined with the ability of this measure to reduce the likelihood of users encountering CSAM mean that costs would need to be particularly high to mean that the measure becomes disproportionate.
  - b) There would be a high risk of displacement of users to smaller services that did not implement the measure as the barriers to switching to a different search service are relatively low. This could significantly reduce the effectiveness of the measure if it were only applied to large general search services.
  - c) Most of the costs are one-off costs and we are aware of relatively small search services that already apply this measure. This indicates that services do not have to be large to implement it successfully.
- 15.81 We are therefore proposing to recommend in our CSEA Code of Practice that all general search services should ensure that URLs known to contain CSAM, or which are hosted by a domain that exists predominantly or entirely for the distribution of CSAM, are deindexed. Services can comply with this by implementing the change if they control the index, or by contracting on a basis which secures that their index supplier implements the change, if they do not control the index.
- 15.82 As discussed above, to ensure the effectiveness of the measure, they should:
- a) make use of an appropriate list of CSAM URLs sourced from a person with expertise in the identification of CSAM and who has arrangements in place to:
    - i) identify URLs suspected to be a CSAM URLs;
    - ii) secure (so far as possible) that CSAM URLs are correctly identified before they are added to the list;
    - iii) regularly update the list with CSAM URLs;
    - iv) review CSAM URLs on the list, and remove any which no longer contain CSAM; and
    - v) secure the list from unauthorised access, interference or security compromises through attacks by bad actor (whether by persons who work for that person, or by any other person).
  - b) Regularly monitor the relevant list to identify new CSAM URLs and reverse deindexing for any URL that is subsequently removed from the list.



- c) Take appropriate steps to secure any internal record of the deindexed CSAM URLs from unauthorised access, interference or security compromises through attacks by bad actor.
- 15.83 We consider that this measure would help general search services meet their safety duty under section 27(3), under which search services must operate the service using proportionate systems and processes designed to minimise the risk of individuals encountering search content that is priority illegal content or other illegal content that the provider knows about. Removing the offending URL, whether other instances of the content remain discoverable, would support compliance with the safety duty.
- 15.84 Vertical search services do not index URLs, and there is a lack of any evidence to suggest that vertical search services play a role or would be likely to play a role in the dissemination of priority illegal content or other illegal content. We therefore do not propose that this measure should apply to them.

## Other policy options for URL detection

- 15.85 We are proposing to take the more prescriptive approach to deindexing CSAM URLs, which we set out above, because we have a robust evidence base on CSAM URL detection. Our evidence regarding the accuracy and efficacy of URL blocking as a tool to combat other illegal harms, such as terrorism, is more limited. As we set out in Chapter 13, we are therefore proposing to allow services more discretion about the circumstances in which they deindex other types of illegal content and the circumstances in which they downrank it.

## Keyword detection regarding articles for use in frauds

---

- 15.86 As explained in Chapter 14, 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.<sup>323</sup> Content amounting to any of these offences is therefore recognised by Schedule 7 of the Act as priority illegal content, and we refer to it in this section as content amounting to an offence concerning articles for use in frauds.<sup>324</sup>
- 15.87 We have explained in Chapter 14 that we are proposing to recommend that some regulated user-to-user services apply standard keyword detection technology to identify content that is likely to amount to these offences (and that such content should be considered by services in accordance with their internal moderation policies).
- 15.88 We have considered the case for recommending that search services use keyword detection technology to reduce the risk of users encountering such content in response to search requests.
- 15.89 In principle, we think that such a measure could be effective. However, for the reasons set out below, we are not proposing to recommend this in our first Code of Practice. We are however keen to gather evidence on this from stakeholders.

---

<sup>323</sup> In Scotland, this is covered by a similar but separate offence - Section 49(3) of Criminal Justice and Licensing (Scotland) Act 2010.

<sup>324</sup> As explained in Ofcom's draft Illegal Content Judgment Guidance, content online is most likely to be 'offering to supply' articles for use in frauds.

## Harms this measure would seek to address

- 15.90 Articles for use in frauds can include fraud ‘guidebooks’ (which provide tips to perpetrators on how to commit fraud), and information on how to access stolen personal and financial credentials.
- 15.91 Research completed by Ofcom has found that content offering to supply articles for use in frauds is easy to find and prevalent in or via search results on some search services. Search queries used in this research returned large volumes of content within the first 20 search results, which we categorised as ‘likely to be prohibited’.<sup>325</sup> We found that search services direct users to ‘likely to be prohibited’ content on U2U services.
- 15.92 This research also generated a range of insights around the use of specific terms associated with the sale of stolen credentials online. Notably, the research found that slang, coded language and more detailed search keywords or queries led to a higher proportion of content likely to be considered as supplying articles for use in frauds. Community-specific language (e.g., slang terms like “Fullz”) was particularly effective at surfacing content on some search services. This indicates that while such content was generally accessible to a user, it was especially so if a user was familiar with relevant terminology.<sup>326</sup> It is for this reason that content of this nature is less likely than other illegal content to come to a service’s attention through standard user reporting. The use of keyword detection technology could therefore in principle provide a means to bolster the search service’s ability to detect content of this nature.
- 15.93 The research also found that several searches using fraud-specific terminology returned links to the dark web within the first 20 results. Beyond the risks of facilitating this priority offence, these sites pose risks to users and their devices, as well as raise serious concerns about potential exposure to further illegal activity due to the higher rate of illegal and malicious activity in such decentralised online spaces. Further evidence suggests that there is a clear link between the supply of articles for use in frauds on the dark web and search services.<sup>327 328</sup> Research commissioned by Cifas aligns with our own research, highlighting the prevalence of this type of content online and the use of specific terms associated with articles for use in frauds.<sup>329</sup> We set out more information about the harm this type of content can cause in chapter 14.
- 15.94 A measure that enables search services to identify search content that amounts to a priority offence regarding articles for use in frauds (and to take appropriate action such as deindexing or downranking that content) would help disrupt fraud activities, and reduce harm to individuals. Specifically, it would:

---

<sup>325</sup> ‘Likely to be prohibited’ is a term developed specifically for the purposes of Ofcom’s research. A full explanation of this can be found in section 3.3 ‘Assessment of search results and webpages’ of that research.

<sup>326</sup> Ofcom, 2023. [Online Content for use in the commission of fraud -accessibility via search services. 18 September 2023](#) [accessed 18 September 2023].

<sup>327</sup> Alex Hern, 201. [The Guardian: Stolen credit card details available for £1 each online](#) [Accessed 13 September 2023].

<sup>328</sup> Paul Bischoff, 2023. [Dark web prices for stolen PayPal accounts up, credit cards down: report](#) [Accessed 13 September 2023].

<sup>329</sup> Cifas and Forensic Pathways, 2018. [Wolves of the Internet: Where do fraudsters hunt for data online](#) [Accessed 13 September 2023].

- a) make it harder for fraudsters to market the proceeds of criminal activity, and therefore diminish the attractiveness of those original illegal activities (i.e., the theft of personal details),
- b) make discoverability of open-source and dark-web sites offering this content more difficult, ultimately reducing the ability to commit fraud using the credentials or guidebooks that are available on websites indexed on the search engine,
- c) limit easy surface web access to sources of stolen financial credentials, which means that opportunistic fraudsters will be disincentivised, and
- d) protect users from becoming victims of fraud, resulting in less financial and emotional distress.

## Options considered

15.95 We have considered the case for recommending in our Code of Practice that search services use keyword detection technology to reduce the risk of users encountering, in response to search requests, search content<sup>330</sup> which amounts to an offence concerning articles for use in frauds.

15.96 We recognise that there may be a range of ways in which services could in principle do this. It might involve the use of standard or more advanced keyword detection technology.<sup>331</sup> The options could include:

- a) Recommending the deployment of keyword detection technology at a relevant point in the search indexing phase and / or ranking phase.
- b) Recommending the deployment of keyword detection technology at a relevant point on the user-facing side (i.e., when a user inputs search queries). This could be used, for example, to prevent users from receiving any results if it is clear from their search query that they are looking for illegal content.

15.97 However, we have limited knowledge around how keyword detection technology might be effectively deployed by different search services, particularly in the context of automated indexing and ranking processes. This is due to the varying technologies required for indexing and ranking, making it challenging to ascertain the most suitable technology for different types of search services. Further, while we understand that keyword detection technology is an established technique for both user-to-user and search services, used in information retrieval and data analysis, there is less evidence of its specific utilisation by search services to deprioritise or deindex illegal or violative content.

15.98 We also currently do not have sufficient information to assess the potential impact on users' rights, particularly on freedom of expression and impacts on persons that rely on search services to advertise legitimate businesses, including SMEs.

15.99 In light of the above, we are currently not in a position to make a recommendation. However, we are keen to receive stakeholders' views (and accompanying evidence) on the ways in which search services might be able to effectively and proportionately deploy

---

<sup>330</sup> This would not include paid-for advertisements.

<sup>331</sup> In the case of keyword detection at the point that users input search queries, this could for example be to prevent users from obtaining any results (i.e., where it is clear from their search request that they are seeking illegal content relating to articles for use in frauds), to scan any search content before it is provided to the user.

keyword detection technology (or other measures that they might be able to take) to reduce the risk of users' encountering search content that amounts to an offence concerning articles for use in fraud.

- 15.100 This includes evidence on the accuracy, effectiveness and lack of bias of those technologies, on the costs associated with this (including the potential scale of returns), and on any safeguards that might be needed to protect users' rights.

# 16. Reporting and complaints

## What is this chapter about?

The Act requires that all U2U and search services must:

- **Have easy to use complaints process, which allow for users to make complaints**, such as: complaints about the presence of illegal content; appeals where content may have been incorrectly identified as illegal; complaints about reporting function; complaints about a service not complying with its duties; complaints about the use of proactive technology in a way that is inconsistent with published terms of service; and
- **take appropriate action in response to complaints.**

This chapter sets out the steps we are proposing to recommend for services to comply with these duties and includes our reasoning and supporting evidence for our proposals.

## What are we proposing?

We are making the following proposals for all U2U and search services:

- **Have complaints processes which enable UK users, affected persons and (for search services where relevant) interested persons, to make, for example, each of the types of complaint highlighted above.**
- **Have an easy to find, easy to access and easy to use complaints system** including: easily findable and accessible content reporting tools and ways to make other kinds of complaint; as few steps as reasonably practicable to make a complaint; ability for UK users to provide context/supporting material; and information and processes to be accessible and comprehensible, including having regard to users with particular accessibility needs such as children (if children use the service) and those with disabilities.
- **Acknowledge receipt of each relevant complaint with indicative timeframes for deciding the complaint.**
- **Actions services should take in response to each type of complaint**, such as: (a) where there are reasonable grounds to infer that content is illegal, U2U services should take this down; (b) illegal content complaints should be handled in accordance with our proposed content moderation and search moderation recommendations; (c) where an appeal is successful, the complainant's content and/or account should be returned to their original position – for example, if content has been erroneously taken down on the basis that it was incorrectly judged to be illegal, or an account banned or suspended erroneously, they should be reinstated, and if a search engine has erroneously downranked or deindexed a webpage on the basis that it was incorrectly judged to contain illegal content this should be reversed.

We are making the following proposals for all large services with a medium or high risk of fraud:

- **Establish and maintain a dedicated reporting channel for fraud, for trusted flaggers.** Within this recommendation, a 'trusted flagger' is each of the following: HM Revenue and Customs (HMRC), Department for Work and Pensions (DWP), City of London Police (ColP), National Crime Agency (NCA), National Cyber Security Centre (NCSC), Dedicated Card Payment Crime Unit (DCPCU), and the Financial Conduct Authority (FCA). This is to enable better

engagement between expert third parties with the competence, expertise and knowledge to detect and investigate fraud (including relevant law enforcement, government departments and regulators), and online services.

## Why are we proposing this?

Complaints are important mechanisms for services to become aware of harmful content. Our proposals are designed to ensure that reporting and complaints functions operate effectively. We consider this will make services better able to identify and remove illegal content, thereby reducing harms to users.

Dedicated reporting channels provide an easy way for expert ‘trusted flaggers’ to report problems to platforms. 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 services and that the creation of a dedicated reporting channel would play an important role in addressing this problem.

## What input do we want from stakeholders?

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

## Introduction

---

- 16.1 Enabling users of regulated services to make certain types of complaints can help in ensuring services are safe by design, accountable and respect users’ rights.<sup>332</sup> This includes enabling users to complain about illegal content in the UK, about their content being taken down or demoted, about their access to a service being restricted based on content moderation decisions, or about infringements of the illegal content safety duty and other relevant matters.
- 16.2 The Act places requirements on services relating to all of these types of complaints. It also contains additional requirements relating to the way in which services must enable users and affected persons to report illegal content. Generally, in this chapter we will refer to all types of reports and other complaints as ‘complaints’. Where necessary, we will distinguish between reports and other complaints when setting out our proposed recommended steps for how services might need to respond to these.
- 16.3 Complaints processes can highlight potentially illegal or other violative content that has been previously undetected by content moderation systems. They provide users with a way to make services aware of this content and for services to take appropriate action, such as swift removal (or in the case of search services, de-indexing or downranking). This reduces the risk of other users encountering illegal content. Our evidence shows that while many services provide content reporting tools or complaints functions, these are not always provided in a way that is accessible, easy to use and transparent. This can act as a barrier to complaints being made.

---

<sup>332</sup> See section 1(3) of the Act.

- 16.4 We recognise that not all complaints can be dealt with immediately, for example if they are complex complaints requiring investigation, and that large numbers of complaints about the same piece of content may be more efficiently dealt with together as one complaint.
- 16.5 Monitoring of complaints data and trends over time may be used by service providers to inform improvements to the systems and processes that a service operates (such as content moderation, the design of recommender systems, or reporting and complaints processes).
- 16.6 In the context of reporting illegal content on services, we have explored the use of dedicated reporting channels (DRCs), as a way to highlight with a greater degree of expert knowledge the presence of illegal content on a given service. A DRC is a means for reporting problems, for example an inbox, a web portal or another relevant mechanism for reporting. The users of DRCs are often referred to as reporters or trusted flaggers. Trusted flaggers are typically entities, and not individual users, that have particular expertise and competence for the purposes of detecting, identifying and notifying services about illegal content. As set out below, we propose a recommendation for large services, with a medium or high risk of fraud, to operate a DRC in relation to fraudulent content. We will continue to develop our understanding of DRCs over time and are likely to return our attention to their possible use for other harms in future consultations.
- 16.7 The Act also requires services to make the policies and processes that govern the handling and resolution of complaints publicly available and easily accessible (including to children). We cover this in our measures on terms of service in Chapter 17.
- 16.8 Services should be aware that further proposals on complaints may be made relating to duties concerning the protection of children, and for those services designated as Category 1 providers, to be set out in future consultations.
- 16.9 It is also important to note key interdependencies and links between the operation of the proposed measures in this chapter and proposed measures set out in other chapters to this consultation, as well as Ofcom guidance, notably:
- a) Governance and accountability;
  - b) Content moderation and Search moderation;
  - c) Terms of Service and Publicly Available Statements; and
  - d) Ofcom’s Illegal Content Judgements Guidance.
- 16.10 The remainder of this chapter sets out the proposed recommended measures for services to comply with their reporting and other complaints duties in the Act. It includes the rationale for the measures we are proposing to recommend, options considered or rejected, our assessment of cost and rights impacts, and our proposed recommendations.

## **Measure 1: All user-to-user and search services must enable users to make relevant complaints**

---

### **Harms that the measure seeks to address**

- 16.11 Sections 21(2)(a) and 32(2)(a) of the Act require that all user-to-user and search services provide a means for users to submit relevant kinds of complaints.
- 16.12 The types of complaints the Act expressly requires services to process are slightly different between U2U and search services. They are referred to as ‘relevant’ complaints.



16.13 For U2U services they are:

- a) Complaints by users and affected persons about the presence of illegal content (this might relate to individual items, or to systemic issues where illegal content is perceived to be prevalent on a service or parts of that service).
- b) Appeals by users whose content may have been incorrectly identified as being illegal, leading to removal or demotion of the content, or to restrictions on users, including warnings, suspensions and bans.
- c) Complaints relating to the operation of an effective reporting function (this might cover, for example, complaints about technical issues preventing users reporting content they consider to be illegal).
- d) A service not complying with its duties about illegal content, or its freedom of expression and privacy duties.
- e) The use of proactive technology leading to the takedown, restriction or deprioritisation of content in a way that is inconsistent with published terms of service. These kinds of complaints must be dealt with by the service regardless of the reason for the action – it need not be that the content is considered to be illegal.

16.14 For search services, relevant complaints are:

- a) Complaints by users and affected persons about search content (ie the results of a user search query) which they consider to be illegal content.
- b) A service not complying with its duties about preventing users encountering illegal content, or its freedom of expression and privacy duties.
- c) Operating an effective reporting function (this might cover, for example, complaints about technical issues preventing users reporting search content they consider to be illegal).
- d) Appeals by an interested person whose website or database may have been incorrectly identified as containing illegal content, leading to its being removed from or demoted in search results.
- e) The use of proactive technology leading to the removal or downranking of search content in a way that is inconsistent with a search service's published terms of service. As above, these kinds of complaints must be dealt with by the service regardless of the reason for the downranking – it need not be that the content is considered to be illegal.

16.15 Although the Act contains separate provisions for reporting and complaints functions, a report is a type of complaint. Services could operate a combined reporting and complaints function for most users and most types of complaints, rather than having two separate user facing processes, as long as it was clear how to use it in each scenario it must cover.

16.16 However, we are of the view that a service cannot use its content reporting tool to receive all relevant complaints, because it needs to be able to receive complaints about the effective operation of the content reporting tool itself. It follows that there must be at least one other means for users to communicate with the service, besides its usual reporting tool.

16.17 Section 227(2) of the Act provides that to be a 'user', it does not matter whether the person is registered to use a service. Therefore, all UK based users and affected persons who can view, or are affected by, content on a service must be able to report or complain about that content, regardless of whether they are registered with that service.

- 16.18 Although the Act requires services to make it possible for relevant complaints to be made and for appropriate action to be taken in relation to them, it does not require services (necessarily) to tell users or URL owners when action has been taken for a reason which would give them a right to have their complaint considered.
- 16.19 It appears to us that to comply with the Act, services would need to do at least one of the following:
- a) make users and website owners<sup>333</sup> aware of when they are entitled to have an appeal considered, which would involve telling them when their content was treated in a particular way because of an illegal content judgement; or
  - b) be able to recognise appeals from users and website owners who have that right, even if the user or website owner concerned did not know it themselves; or
  - c) handle all complaints as if the duty applied to them.
- 16.20 As set out in paragraph 12.40 of the U2U content moderation chapter, many services will have designed their terms of service or community guidelines to comply with laws in multiple jurisdictions in which they operate or where their services are targeted. The Act provides for services to have different terms of service for UK users when compared to users elsewhere in the world. In practice, where the Act requires illegal content to be taken down, this means taken down for UK users.
- 16.21 We think that in the first instance, this means services have a choice. They may choose to do all the things the Act requires, for all their users no matter where in the world they are located. But they may instead choose to do those things only in relation to their UK users.
- 16.22 If a service wishes to comply with its duties around reporting and complaints in this narrower way, it will first of all need to know if the user who has submitted an illegal content complaint has been served this content in the UK.
- 16.23 Search services may also need a way for complainants to tell them if they are an 'interested person', because 'interested persons' are the only ones with a right to make certain kinds of complaints. The Act defines an interested person in relation to a search service as *"a person 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 UK; (b) in the case of an entity, the entity is incorporated or formed under the law of any part of the UK"*.

## Options and effectiveness

- 16.24 We considered whether to recommend that services notify users and website owners when they are entitled to have an appeal considered.
- 16.25 However, we felt this may unduly constrain services in how they ensure that the complaints they must consider are dealt with, consistent with data protection laws and other

---

<sup>333</sup> The right to have an appeal considered arises for 'interested persons', defined, in relation to a search service or a combined 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; or (b) in the case of an entity, the entity is incorporated or formed under the law of any part of the United Kingdom. For readability, we have used the term 'website owners' as shorthand to refer to interested persons in this section.

obligations.<sup>334</sup> We also think it likely to be impossible for search services to do this, as they often do not have a direct relationship with website owners.

- 16.26 We also considered whether we should recommend that services keep records to enable them to identify whether a user had an appeal right. However, we considered that this is likely to be unduly onerous given that some services may well handle all complaints. We noted that if a service wishes to handle all complaints, but only from UK users, one way of achieving this might be for a service to provide, as part of its complaints process, a way for users to indicate in which jurisdiction they are filing a report or complaint.
- 16.27 We therefore propose to recommend in relation to all user-to-user and search services, that their complaints processes enable UK users, affected persons and (for search services where relevant) interested persons respectively, to make each type of relevant complaint in a way which will ensure that the service will take appropriate action in relation to them. We consider this the minimum necessary to comply with the Act.

## Costs and risks

- 16.28 Handling relevant complaints is required by the Act. Given that our recommendation closely follows the specific requirements in the Act, and leaves the widest possible discretion to services on how to achieve what is required, we consider its impacts are as required by the Act.
- 16.29 Services can decide the most appropriate and proportionate approach for their own contexts, and the set up and operating costs that flow from that are costs imposed by the Act. This flexibility will allow them to take an approach proportionate to the risks they carry.

## Rights impacts

- 16.30 We do not consider that there are any freedom of expression or freedom of association impacts in relation to this proposed measure.
- 16.31 This proposal may involve some data collection as it relates to content or user access decisions made by the service on the basis of illegal content being shared or published. However, if a service decides that it will maintain records of such decisions, it will need to do so consistently with its duties under data protection and privacy laws.

## Provisional conclusion

- 16.32 For all the reasons above, we propose to recommend that, in relation to all user-to-user and search services, they have complaints processes which enable UK users, affected persons and (for search services where relevant) interested persons, to make each type of relevant complaint in a way which will secure that the service will take appropriate action in relation to them.
- 16.33 The handling of complaints is required by sections 21 and 32 of the Act, which are not part of the safety duty, and so this proposed measure belongs in our Code on other duties. However, we consider that handling complaints about illegal content is also necessary for a service to meet its safety duties in relation to CSEA and terrorism content<sup>335</sup>, and therefore

---

<sup>335</sup> Specifically, for U2U services, their duties relating to minimising the length of time for which any priority illegal content is present; and where the provider is alerted by a person to the presence of any illegal content,

also propose to also include the provision relating to those sorts of complaints in our CSEA and terrorism Codes.

## Measure 2: All search and user-to-user services must provide an easy to find, easy to access and easy to use complaints system

---

### Harms that the measure seeks to address

- 16.34 Sections 20(2) and 31(2) of the Act place duties on user-to-user and search services to operate systems and processes that allow users and affected persons in the UK to ‘easily’ report illegal content.
- 16.35 Complaints processes for all types of relevant complaint must be easy to access, easy to use (including by children) and transparent (see section 21(2)(c) and 32(2)(c) of the Act).
- 16.36 Users will differ in what they find ‘easy’ when reporting or complaining, depending on many factors such as age, media literacy, cognitive limitations, learning difficulties and access needs such as visual impairment. There is a risk of illegal content remaining on a platform for longer if users struggle to make a report, increasing the likelihood of harm to more users who may then come across it or be subject to repeated exposure. There are related risks when other complaint types are considered, such as those related to operating an effective reporting system, or complaints related to user rights. In all of these cases, we believe an easy to use and accessible process for reports and other complaints would reduce risks in these areas.
- 16.37 If complaints systems are difficult to find or not clearly identifiable as such, users and affected persons<sup>336</sup> may not be able to locate or access them, making it hard to flag potential illegal content or other problems. There is therefore a risk that some users will give up trying to complain, leading to illegal content being available on services for longer periods, or other relevant matters going unchecked potentially creating risks of harm to users.
- 16.38 Evidence from our own research<sup>337</sup> and concerns expressed in responses to our 2022 Illegal Harms Call for Evidence<sup>338</sup> suggest that the following issues may make the use of complaints processes particularly difficult for some users:
- a) Users may not have the ability or time to read lengthy complicated information, so complexity reduces the likelihood of complaining;
  - b) Without a recognisable reporting icon, for example a flag, users may struggle to know how to begin filing a complaint;

---

swiftly take down such content (section 10(3)). For search services, their duties relating to minimising the risk of individuals encountering search content which is priority illegal content (section 27(3)).

<sup>336</sup> See section 74(6)

<sup>337</sup>, Ofcom 2022. [Video-sharing platform users' experiences and attitudes report](#), Ofcom 2023. [Behavioural insights research - online safety: understanding the impact of video sharing platform \(VSP\) design on user behaviour](#), [Online Nation Report](#), Ofcom March 2023 [Children and Parents: Media Use and Attitudes 2023](#).

<sup>338</sup> Ofcom [2022 Illegal Harms Call for Evidence](#) (Evidence from specific stakeholders is included from para 16.45 below)

- c) They may find the complaints process too onerous if there are too many steps to report or submit a complaint.
- 16.39 If users cannot add context to their reports or complaints, the content being complained about could seem harmless or innocuous to a content moderator, and so not be acted on appropriately. This can include cases of domestic violence or harassment (for example, the posting of an image of someone's front door which by itself and without context would not appear to be harmful). Ofcom's draft Illegal Content Judgments Guidance provides further information on how and when services may need to take contextual information into account in considering whether content is illegal content.
- 16.40 For other types of complaints, for example those related appeals against action taken against the user or their content because their content is thought to be illegal, we think that providing users with the ability to furnish service providers with contextual information related to their complaint would support ease of use and accessibility requirements.
- 16.41 Users may be better served by the ability to make a single report covering multiple cases of illegal content (such as for persistent abuse) rather than separate reports for each post which could be time consuming, discourage reporting or not allow moderators to see the extent of the harm or relevant matter being reported or complained about.
- 16.42 In our view, contextual information is also important when reports or complaints about illegal content from a UK perspective are considered.

## Options considered

- 16.43 In considering an approach to making reporting and complaints systems easy to access and use for different user groups, we considered two high level options, informed by stakeholder comments:
- a) **Option 1:** recommending specific design features or requirements in services' reporting and complaints systems including:
    - i) Specifying what a reporting tool must look like (e.g. a flag icon);
    - ii) Requiring services to provide set categories for complaints;
    - iii) Providing reporting or complaints processes in specific languages;
    - iv) Active prompts to encourage complaints.
  - b) **Option 2 (recommended):** setting out the high level requirements that would secure compliance with the relevant duties, but not setting out exactly how services should design their reporting and complaints systems. These requirements would include:
    - i) the accessibility and findability of both content reporting tools and the way to make other complaints;
    - ii) the number of steps needed to complain;
    - iii) the ability for users to include relevant context when submitting a complaint;
    - iv) accessibility and comprehensibility of information relating to the complaints processes, having regard to the findings of their risk assessment in relation to the accessibility needs of their userbase.

## Discussion of options

- 16.44 When assessing these options we considered what might be effective and proportionate bearing in mind the number, variety and sizes of services within the Act's scope.

- 16.45 An effective and proportionate reporting or complaints system depends to some degree on the specific context of the service. For example, in response to our 2022 Illegal Harms Call for Evidence, the Federation of Small Businesses said it is imperative that any measures that businesses will have to take are proportionate to the risk of harm, and will not have a disproportionate impact on their ability to innovate and compete. It warned against a ‘one-size-fits-all’ approach to online safety regulation<sup>339</sup>.
- 16.46 We therefore do not think it would be appropriate to define the specific features that reporting or complaints processes should have at this stage. We have limited evidence to inform our understanding of the efficacy and costs of implementing specific features. We also believe that this approach may create risks in terms of disincentivising innovation, prompting services to incur potentially disproportionate costs, and unintended consequences given our limited understanding of the particular circumstances and risks for any given service.
- 16.47 On this basis, we think option 2 (defining high level considerations in the design of reporting and complaints functions) strikes a better balance at this stage between ensuring that services which adopt our recommended measures are in compliance with their duties, while also enabling services to do this in a way that is most appropriate and proportionate for their context. In particular:
- In relation to reporting tools, we consider that if these tools are easily accessible and clear, it is not necessary to prescribe exactly what they look like.
  - While we acknowledge that setting out different categories of complaints can be useful, and we would therefore want to encourage services to include these where appropriate, we think that services should set these out according to their particular user base, risk profile and terms of service. Many of the largest and riskiest services already have categories to help users with submitting their report or complaint but have implemented this in different ways. For example, some may prefer to combine illegality with other violative categories where many users may not easily be able to tell the difference, or to focus on the main types of issue on the service, so as to make it simpler for users to report.
- 16.48 We also considered recommending that UK users should be able to complain in specified languages, as content moderation systems may not recognise multiple languages and dialects, which may lead to illegal content remaining online. However, as set out paragraph 12.164 of the content moderation chapter, the language expertise required to deal with the risk of harm in a particular language will likely differ from service to service based a number of factors, including user base, content type and functionality. For this reason, we feel Codes should not be prescriptive around what exact languages complaints processes should cater for on a particular service. Instead, we propose to specify in our proposed measure that services should have regard to their user base and risk profiles when considering accessibility.
- 16.49 Finally, we have some evidence on the application of active prompts to assist users with reporting or complaining: our own behavioural trial research indicates they can encourage reporting. We consider a particular proposal which may prompt children to make complaints as a part of a suite of proposals relating to grooming, in the section entitled Support for child users in Chapter 18 (U2U default settings and support for child users). This is because we

---

<sup>339</sup> [Federation of Small Businesses response to Ofcom 2022 Illegal Harms Call for Evidence](#) p1

have clear evidence that children do not always make appropriate complaints. Whether to make wider recommendations on prompts is an area we would like to keep under review and do further work on.<sup>340</sup>

#### *Discussion of our proposed recommendation – option 2*

- 16.50 The Act requires that all services should make it easy for users and affected persons to make complaints, including in particular reports about suspected illegal content.
- 16.51 We consider that affected persons and unregistered users should be able to complain using the same processes as registered users. While we have less information about how affected persons may interact with complaints processes, the Act requires services to consider their needs. We believe that, if services designed their complaints processes to be accessible to all users – including children and non-registered users – this would make them easier to understand and use by everyone.
- 16.52 In particular, children appear to benefit from easy-to-use systems. Although in some instances parents or other adults will make complaints on behalf of children as ‘affected persons’, children and young people should not be excluded from the opportunity to make use of these facilities themselves, particularly in relation to illegal content. It is important that children are not dissuaded from pursuing complaints due to shortcomings in the design of a providers’ procedures about how to make relevant complaints, whether due to technical language being used, or making access to the complaints process unnecessarily difficult.
- 16.53 Our evidence set out below indicates that the risks of illegal content being widely disseminated can be reduced by services providing clear and accessible reporting and complaints procedures, although this may vary depending on what steps a service is taking to identify such content proactively itself.
- 16.54 We have based our proposals on a range of evidence, including our own research (including online behavioural insight trials on reporting features)<sup>341</sup>, stakeholder responses to our 2022 Illegal Harms Call for Evidence<sup>342</sup>, government guidelines<sup>343</sup> and civil society reports.<sup>344</sup> From these sources, we have identified the following key design elements which we provisionally think are essential to make it easy for users to report or complain, while allowing sufficient flexibility for services to design their systems in a manner that can be most effectively and proportionately applied to their platforms. We will step through each of these elements and the rationale in turn.

---

<sup>340</sup> Our [Behavioural insights research](#) found that by including an active prompt to report when users chose to dislike or comment on a video increased reporting significantly. This is because some users use the dislike or comment functions as a way to show their displeasure, and a timely prompt to report content helped to translate that displeasure into a submitted report. We have also conducted [further behavioural research](#) which looked at other types of interventions to help boost users’ capability of reporting.

<sup>341</sup> Ofcom 2023. [Behavioural insights research - understanding the impact of video sharing platform \(VSP\) design on user behaviour](#).

<sup>342</sup> Ofcom [2022 Illegal Harms Call for Evidence](#)

<sup>343</sup> DSIT and DCMS 2021. [Child online safety: Age-appropriate content](#)

<sup>344</sup> Rights for Children. [Your right to complain](#), [accessed 8 September 2023]; Centre for Countering Digital Hate. [STAR Framework – CCDH’s Global Standard for Regulating Social Media](#) [accessed 8 September 2023]. Subsequent references throughout; Carnegie UK, The End Violence Against Women Coalition, Glitch, NSPCC, Refuge, 5Rights, Woods, L., and McGlynn, C., 2022. [VAWG Code of Practice](#), [accessed 8 September 2023].



16.55 We now move on to discuss in turn each of the high level requirements related to accessibility of complaints systems which we are considering for inclusion in Codes.

## Measure 2 (a) – illegal content reporting functions or tools should be easy to find in relation to the content being viewed and easily accessible; and the way to make other complaints should be easy to find and easily accessible

16.56 Our research suggests that where content reporting tools are hard to find this can act as a deterrent to users who would otherwise wish to make a content-related complaint, including complaints about suspected illegal content. For example:

- Our VSP tracker published in September 2023<sup>345</sup> found that of those who claim to have been exposed to perceived harmful content on VSPs, 14% said they tried to use a reporting mechanism but could not because it was too hard to find.
- Our research into the impact of behaviourally informed designs for content-reporting mechanisms for VSPs<sup>346</sup> found that making the reporting function more prominent led to a statistically significant increase in the number of users reporting content they were concerned about.<sup>347</sup>

16.57 The *Centre for Countering Digital Hate* said in its response to our 2022 Illegal Harms Call for Evidence that it found in some cases it was difficult for users to find a reporting system and recommended that “platforms can improve how to find a reporting function by adopting a safety by design approach”.<sup>348</sup> Others pointed to easy to use reporting functions such as clickable buttons or functions being made clear to users by using an easily recognisable symbol, such as a flag, or words like ‘Report’.<sup>349</sup>

16.58 We also note precedent for the highest risk services in the Australian Social Media Online Safety Code, which states that “providers of a social media service should ensure that reporting tools are integrated within the functionality of the social media service in a manner that is visible and accessible at the point the Australian end-user accesses materials posted by other end-users”.<sup>350</sup>

16.59 We do not consider that a report can be made ‘easily’ if it is not clear to users where and how they can report content they consider to be illegal. Reporting functions or tools should be easy to find and easily accessible in relation to the content being viewed.

16.60 We have less evidence about the difficulty of making other kinds of complaints being a problem for users. However, the Act requires those kinds of complaints procedures to be ‘easy to use’. We do not consider that a complaints procedure would be easy to use if it was

---

<sup>345</sup> Ofcom 2023 [VSP tracker](#)- Video-sharing platform users' experiences and attitudes

<sup>346</sup> Ofcom 2023 [Behavioural insights for online safety: understanding the impact of video sharing platform \(VSP\) design on user behaviour](#), p6.

<sup>347</sup> We found that the percentage of participants reporting one potentially harmful video increased from 1% in the control group to 4% in the treatment group in which the ellipsis was replaced with a flag icon – a fourfold increase.

<sup>348</sup> Centre for Countering Digital Hate, 2022

<sup>349</sup> Children and Tech said in our 2022 Illegal Harms Call for Evidence “Create easy to use reporting functions (such as clickable buttons) and to have clarity in where to go to report.....”. Source: Children and Tech response to Ofcom 2022 Illegal Harms Call for Evidence, subsequent references throughout; [TrustElevate](#) response to Ofcom 2022 Illegal Harms Call for Evidence Q7 p8, subsequent references throughout.

<sup>350</sup> Australian E-Safety Commissioner, 2023 Australian Social Media Online Safety Code, p17.

not clear to users where and how they can make a complaint. We therefore consider that processes for making other kinds of complaints should be easy to find and easily accessible.

### Measure 2(b) – the number of steps necessary (such as the number of clicks or navigation points) for users and affected persons to submit any complaint are as few as is reasonably practicable

- 16.61 Our Behavioural insights research cited above highlighted that having too many steps in a reporting or complaints process made users less likely to engage with such processes, including where the tool itself was hidden behind an ellipsis.<sup>351</sup>
- 16.62 In response to our 2022 Illegal Harms Call for Evidence, TrustElevate recommended that *“reporting mechanisms should include the minimum number of clicks and steps for a user to quickly submit a report or complaint with ease while equipping the receiving party/platform with sufficient information to assess the report and determine the appropriate response”*.<sup>352</sup>
- 16.63 Reducing reporting steps could particularly help those with learning difficulties, for example, who may struggle with processes involving a number of complicated steps.
- 16.64 Given the large number and variety of different types of service, we provisionally consider it would not be appropriate to set out the maximum number of steps required to make a report or another complaint.
- 16.65 However, the number of steps necessary (such as the number of clicks or navigation points) for users and affected persons to submit any complaint should be as few as is reasonably practicable.

### Measure 2(c) – users and affected persons can provide relevant information or supporting material when submitting complaints to a service

- 16.66 There is a risk that if a user is unable to provide supporting information when making complaints, the service may not have the necessary information to make an informed judgement and therefore may decide not to uphold a valid complaint. Individuals may also consider the process difficult to use as a result, and so not complain at all. This could lead to illegal content remaining on the service, systemic issues with reporting or complaining about illegal content or freedom of expression issues not being addressed. It could also lead to other negative impacts on users, for example if they cannot provide relevant information to contest a decision about a restriction placed on their content or account.
- 16.67 Context is often crucial to enable content moderators to correctly identify illegal content. This is supported by multiple respondents to our 2022 Illegal Harms Call for Evidence. For example, the *Antisemitism Policy Trust*<sup>353</sup> cited 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. Without this explanatory context, the photo was deemed not to breach the service’s rules. *Refuge*<sup>354</sup> provided an example of survivors of domestic abuse having received images of their front doors and road signs after moving to a new location. The image of a front door is not harmful in itself and so is unlikely to be removed by content

---

<sup>351</sup> Kantar on behalf of Ofcom [Ofcom online trials: Reporting mechanisms of video sharing platforms](#) p40

<sup>352</sup> TrustElevate, 2022.

<sup>353</sup> [Antisemitism Policy Trust response to Ofcom 2022 Illegal Harms Call for Evidence](#) p3

<sup>354</sup> [Refuge response to Ofcom 2022 Illegal Harms Call for Evidence](#) p3. Subsequent references throughout.

moderators. However, with added context it may be reasonable to infer that the content amounts to harassment. Our draft Illegal Content Judgments Guidance provides more information and examples of how contextual information from complaints (or appeals) may inform judgments about whether content amounts to particular offences.

- 16.68 Similarly, a user wishing to complain about a problem with a service’s reporting tool or some form of non-compliance with the safety duty may need to be able to attach a screen shot or a description to their complaints.
- 16.69 *Refuge*<sup>355</sup> also said 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 re-traumatising process for survivors”. An ability to provide context, for example, screenshots showing how the user is being subjected to a pattern of behaviour or the identities of the accounts engaging in the behaviour concerned, would reduce this burden on the user concerned. We consider this is likely to be helpful for those who are at risk of harm from harassment and offline violence, many of whom are women and girls.
- 16.70 Therefore, users and affected persons should be able to provide contextual information in support of their complaint, including in particular those made using the reporting tool, to help services to determine the complaint and take appropriate action.

## Measure 2(d) – information and processes should be accessible and comprehensible, including having regard to the findings of their risk assessment in relation to the accessibility needs of their UK userbase

- 16.71 The Act contains specific requirements, considered in Chapter 17, about how complaints processes should be described in services’ terms of service and publicly available statements respectively. This section does not consider these.
- 16.72 However, in the course of providing a complaints process, a service will produce other information (for example, the actual reporting forms). The complaints process needs to be accessible in itself, as required by the Act.
- 16.73 We consider there is clear evidence that not all reporting processes for content of concern are sufficiently accessible for vulnerable groups, which deters them from reporting. For example:
- a) Our research suggests that many children do not know how to use reporting systems or find them difficult. Awareness levels for this type of function (35%) were lower than for other types of protective measure, such as blocking people on social media (84%).<sup>356</sup> In our VSP Tracker 2021/2022, 41% of parents called for access to and use of reporting to be made easier for children.<sup>357</sup> A member of 5Rights’ youth advisory group said they have stopped reporting harmful activity because the process is too onerous.<sup>358</sup> We

---

<sup>355</sup> [Refuge, 2022](#). p7

<sup>356</sup> [Ofcom 2023 Children and Parents: Media Use and Attitudes 2023](#) p38

<sup>357</sup> [Ofcom 2022 VSP Tracker 2021/2022 Video-sharing platform users' experiences and attitudes](#), p51

<sup>358</sup> [5Rights](#) response to 2022 Illegal Harms Call for Evidence. Subsequent references throughout; See also [Global Partners Digital](#) response to our 2022 Illegal Harms Call for Evidence, which recommended “more simple and straightforward mechanisms for underage users to lodge complaints, including simpler or more clearly explained categories, simpler language, graphics and visuals to aid explanation and instructions”; [NSPCC](#) response to our 2022 Illegal Harms Call for Evidence, which recommended “providers should make a concerted effort to understand the dynamics of abuse and why children are not using the reporting

provisionally consider that improved reporting accessibility for children would be likely to improve their engagement with the system.

- b) Our Online Experiences Tracker 2021/22 found that users with any limiting or impacting conditions (29%) are more likely to be dissatisfied with the reporting process than the average (25%) and those with no limiting or impacting conditions (21%).<sup>359</sup> Our VSP Tracker published in September 2023<sup>360</sup> discovered that 14% of those with a limiting condition found reporting mechanisms difficult to find compared with 5% of those with no limiting condition.
- c) *Mencap* said that people with a learning disability tell it that routes for raising concerns and registering complaints are often not accessible. It said that “*accessible information, and in particular easy read, should be provided to users by all social media and other providers*”.
- d) As set out in paragraphs 12.158 to 12.164 of the U2U content moderation chapter, language can also be a concern. In 2021, 98.2% of people in England and Wales spoke English (or, in Wales, Welsh) either as their main language or ‘well’. 1.5% could not speak it well, and 0.3%, 161,000 could not speak it at all.<sup>361</sup> However, the particular languages in which users are fluent will likely differ from service to service based a number of factors, including user base, content type and functionality. For this reason, we feel Codes should not be prescriptive around what exact languages complaints processes should cater for on a particular service.

16.74 We provisionally believe that the design of accessible complaints systems which includes accessibility for vulnerable or disabled users would improve outcomes for these users and for all other users and affected persons. We consider that clear and prominent text and icons could help those with vision impairments as well as other vulnerable groups. The *Web Content Accessibility Guidelines* (WCAG), an internationally recognised set of recommendations for improving web accessibility, explain how to make digital services, websites and apps accessible to everyone.<sup>362</sup>

16.75 Although realistically, systems cannot ensure that every user will find it easy to report or complain about content, consideration by a service of all its users’ needs, including children and those with disabilities, should help produce an inclusive system that will be easiest for the largest number of people. Barriers to reporting should be removed as far as possible.

16.76 We therefore provisionally consider that in designing their complaints processes, including their reporting tool or function, services should have regard to the particular needs of their UK user base, including the needs of children (if children use the service) and those with disabilities. Written information should be comprehensible based on the age of the youngest person permitted to agree to the service’s terms of service or publicly available statement; and the process should be designed for the purposes of ensuring usability for those

---

mechanisms. This information should then be translated into redesigning their tools to ensure reporting is more accessible”. See also the Government’s guidance to business [Child online safety: Age-appropriate content](#).

<sup>359</sup> Yonder on behalf of Ofcom [September 2022 Online Experiences Tracker Summary Report 2021/22](#), p17

<sup>360</sup> Ofcom 2023 [VSP Tracker](#) - Video-sharing platform users' experiences and attitudes.

<sup>361</sup> Office of National Statistics, 2022. [Language, England and Wales: Census 2021](#).

<sup>362</sup> UK Government. [Web Content Accessibility Guidelines \(WCAG\)](#).

dependent on assistive technologies including: keyboard navigation; and screen reading technology.

- 16.77 Together with our proposed Measure 1 in Chapter 17 (Terms of Service and Publicly Available Statements), and Measure 3 paragraph 16.117 below on acknowledgements to complainants, this would secure that complaints processes are transparent, as required by the Act.

## Costs and risks

- 16.78 Our proposed measure describes how we recommend services meet the specific requirements in the Act relating to complaints. Given we are not proposing to specify precisely how services should design their complaints systems, and instead propose to set out high level requirements leaving a wide discretion to services on how to achieve what is required, most of the costs of the proposed measure relate to the specific requirements in the Act, over which Ofcom has no discretion.
- 16.79 Services can decide the most appropriate and proportionate approach for their own contexts, and the set up and operating costs that flow from that are costs imposed by the Act. This flexibility will allow them to take an approach proportionate to the risks they carry.
- 16.80 These specific requirements may involve some direct one-off implementation costs for designing required changes, and the engineering costs of testing and implementing those changes. There may also be costs of further refining the complaints system, as services would need to ensure it continues to meet the requirements over time. The level of implementation costs would depend on the complexity of the complaints processes the service adopts. We expect these to vary by service size to some extent, as smaller services will tend to have simpler systems than large services.
- 16.81 There will also be on-going costs of considering complaints. If the complaint process is easier, the volume of complaints is likely to increase, tending to increase costs. This may particularly affect complaints relating to content. To the extent reporting of illegal content increases, then this is the intent of the measure and the costs of dealing with this will tend to increase in proportion to the benefits of the measure.
- 16.82 However, reports may not always be accurate.<sup>363</sup> If reporting is easier, there is also likely to be an increase in the volume of reports where the content is not illegal. Some increase is likely to be inevitable with any measure which meets the requirements of the Act. Therefore, a proportion of any such increase results from the duty on all services in the Act rather than from our specific proposals. However, we recognise that our specific proposals may increase these costs further. For example, our proposal to require the reporting method to be easily accessible in relation to the content in question may increase the amount of reporting of legal content and the costs of handling this. We believe this increase is mitigated through the flexibility allowed in parts of our proposals. In particular, we are not proposing to specify how services should categorise complaints or exactly what services'

---

<sup>363</sup> For example, [TrustPilot's 2021 transparency report](#) says that only 12.4% of consumer user reports in 2021 were deemed to be accurate. [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. 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.

complaints processes should look like. We consider below what appropriate action is in response to complaints.

## Rights impacts

### Impacts on freedom of expression and applicable safeguards

16.83 We did not identify any impacts of this proposed measure on freedom of expression. If anything, enabling users, affected persons and (where relevant website owners) to complain more easily promotes their freedom of expression and helps to safeguard their rights.

### Impacts on privacy and applicable safeguards

16.84 We do not consider that improving the prominence or design of reporting systems would have an impact on privacy. Where reporting and complaints mechanisms involve personal data processing, services must comply with data protection law.

16.85 Asking providers to allow for greater context and information to be provided in support of complaints may engage the right to privacy of users and other affected persons because a greater amount of their own and other people's personal data may be disclosed to the service than would otherwise be the case. However, we think that this is justifiable. First, there is no obligation on complainants to convey personal information if they do not wish to. Second, where any additional personal data is provided, it must be handled in accordance with data protection laws in any event. Finally, the risks are outweighed by the benefits: compliance with this measure should lead to service providers making better decisions pursuant to their illegal content safety duties.

## Provisional conclusion

16.86 Under the Act, regulated user to user and search services are required to have reporting systems and process which allow users and affected persons to easily report content/search content which they consider to be illegal content and complaints processes which are easy to access, easy to use (including by children), and transparent.

16.87 We've set out a number of proposals which we believe, when taken together, would make reporting and complaints processes easy to use and accessible as required by the Act. Whilst the measure will have some costs, we provisionally conclude that given the importance of good reporting and complaints procedures, such costs are proportionate and are primarily based on the requirements of the Act, rather than on regulatory choices made by Ofcom.

16.88 This is particularly the case since: (i) it is difficult to envisage how services could comply with their duties under the Act if they did not follow the measures that we have set out; and (ii) our approach allows services significant flexibility to implement the above measures in a way which is cost effective and practicable for them.

16.89 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice for other duties contain the measures set out above. In relation to search services, references below to 'illegal content' would be replaced with 'search content that is illegal content' and 'interested persons' would be added where appropriate. Please see measure 5B within the draft Code for search services.

16.90 As set out above, the handling of complaints is required by sections 21 and 32 of the Act, which are not part of the safety duty, and so this proposed measure belongs in our Code on other duties. The same is true of the specific requirements relating to reports, in sections 20



and 31 of the Act. However, we consider that handling complaints, including where they are received through a user report, is also necessary for a service to meet its safety duties in relation to CSEA and terrorism content<sup>364</sup>, and therefore also propose to also include this measure in our CSEA and terrorism Codes.

## Measure 3: Sending indicative timelines for considering complaints (U2U and search)

---

### Harms that the measure seeks to address

- 16.91 The Act requires all services to operate processes that provide for appropriate action to be taken in response to reports about illegal content and other types of complaints. One of the key purposes of the Act is to secure that “transparency and accountability are provided in relation to” services,<sup>365</sup> and we consider that what is ‘appropriate action’ by the service provider for the purposes of the complaints handling duties must be considered in that light.
- 16.92 Evidence suggests that complainants can often wait a long time to receive any information about their complaint, and in some cases receive no response at all.<sup>366</sup> Children in particular are often dissuaded from reporting content or complaining as they don’t think anything will come of their complaint.<sup>367</sup>
- 16.93 Some responses to our 2022 Illegal Harms Call for Evidence raised various issues about the way complaints are currently handled by some online services. The inability of complainants to follow up on and check on the status and progress of reports, including by children and disabled users, was raised by some respondents<sup>368</sup> as a barrier to complaining. To address concerns about not being able to follow up, Glitch suggested a specific point of contact should be offered by service providers.<sup>369</sup>
- 16.94 Refuge highlighted in its [Unsocial Spaces Report 2021](#) that in its experience “*a survivor’s priority is often for public abusive content to be removed as quickly as possible. Waiting long periods for a reply to requests for content removal only compounds the stress and trauma they are experiencing*”.<sup>370</sup>

### Options

- 16.95 We identified the following two options for handling communications to complainants:
- a) Complainants receive an acknowledgment of their complaint containing indicative timelines for handling it; or
  - b) A more detailed approach to enable complainants to check the status of their complaints or for updates to be proactively sent to users.

---

<sup>364</sup> Specifically, for U2U services, their duties relating to minimising the length of time for which any priority illegal content is present; and where the provider is alerted by a person to the presence of any illegal content, swiftly take down such content (section 10(3)). For search services, their duties relating to minimising the risk of individuals encountering search content which is priority illegal content (section 27(3)).

<sup>365</sup> Section 1(3)(b)(iii) of the Act.

<sup>366</sup> Refuge, 2021. [Unsocial Spaces Report](#), p7 and 25. Subsequent references throughout

<sup>367</sup> Ofcom 2022. [Online Nation Report 2022](#), p73.

<sup>368</sup> 5Rights 2022; [Mencap](#) response to 2022 Illegal Harms Call for Evidence.

<sup>369</sup> [Glitch response to Ofcom 2022 Illegal Harms Call for Evidence](#). Q10 p5. Subsequent references throughout.

<sup>370</sup> Refuge, 2021. p25



16.96 For the reasons set out below, we provisionally propose to recommend that all services should acknowledge receipt of complaints with an indicative timeframe for deciding the complaint. We do not propose to recommend that services necessarily need to provide a contact person, progress updates or information on the outcome of the complaint.

## Effectiveness

### Acknowledging complaints and providing timeframes

- 16.97 If complainants do not feel that their complaints are being dealt with, there is a risk that they will consider it not worth complaining which could lead to less detection of illegal content. Some respondents to our 2022 Illegal Harms Call for Evidence called for at least an acknowledgement of a complaint within a certain timeframe.<sup>371</sup> *Refuge* suggested 24 hours.<sup>372</sup> Experiences in other sectors show that a response within two working days increases confidence in complaints handling processes.<sup>373</sup>
- 16.98 Several respondents to the 2022 Illegal Harms Call for Evidence<sup>374</sup> supported time frames being provided to complainants. *Trustpilot* told us that “*the first response time to all flagged reviews, in all markets globally, is currently within 48 hours*”.<sup>375</sup> Other respondents suggested it would be helpful to have greater clarity about the process following making a complaint, including timings.<sup>376</sup> Respondents said timescales should be proportionate to the seriousness of a report, which in some instances may require an immediate response.<sup>377</sup> *Refuge* said in its *Unsocial Spaces Report* that complaints about serious offences should be dealt with within 24-48 hours.<sup>378</sup>
- 16.99 Ofcom’s Video Sharing Platform (VSP) framework highlights the importance of setting timeframes for actioning complaints. This can be useful in developing metrics as a way of demonstrating effective procedures for the handling and resolution of complaints.
- 16.100 However, as set out in paragraph 12.110 of the U2U content moderation chapter, we are conscious of the risk of perverse outcomes if the regulator were to suggest a one size fits all approach to deadlines for content moderation processes, including those for complaints. It could lead to resources being diverted from types of illegal content which due to their virality cause harm to very many people, resources being diverted from very serious harms which are not illegal (for example, harms to children), or to decisions being made incorrectly due to time pressures.
- 16.101 We consider that complainants’ concerns about lack of action are likely to be allayed somewhat by having some indicative idea of the timeframe for their complaint to be processed, and that increased trust in the process would have an effect on their likelihood of using it. There is a risk that an indicative timeframe would be misunderstood by

---

<sup>371</sup> Carnegie UK response to Ofcom 2022 Illegal Harms Call for Evidence, Q7 p6 .Subsequent reference throughout; *Glitch* Q7 p5

<sup>372</sup> *Refuge* 2021. p30

<sup>373</sup> Legal Ombudsman [Best practice complaint handling guide](#), paragraph “Complaints Process: Inform.”

<sup>374</sup> *5Rights*, 2022. Q7 p9; Carnegie, 2022. Q7 p6; *Glitch*, 2022. Q7 p5; *Refuge*, 2022 Q7 p6;

<sup>375</sup> *Trustpilot* response to Ofcom 2022 Online Safety Call for Evidence, p18.

<sup>376</sup> *Children and Tech*, 2022.

<sup>377</sup> *5Rights*, 2021. Q7 p9.

<sup>378</sup> *Refuge*, 2021. [p30](#)

complainants as a binding deadline, leading to worse outcomes. However, we consider that services would be able to draft it in such a way that it did not lead to false expectations.

16.102 In addition, a duty to provide indicative timeframes to users would incentivise services to set timeframes which are appropriately swift and to meet them. As set out above, one of the key purposes of the Act is to secure that transparency and accountability are provided in relation to services.

### Enabling complainants to check the status of their complaints or for updates to be proactively sent to users

16.103 Status updates could be provided to complainants e.g. by giving them access to check a database or by sending them updates. Alternatively providing a specific complaints handling point of contact may help facilitate communication between services and users about the status of any complaint, especially in cases when the indicative timeline has not been met or a complaint outcome has not been received. We understand that this is common practice in complaints procedures across many sectors, for example in the postal or telecoms sectors, and could at its simplest level comprise a specific complaints handling email address.

16.104 It appears likely that at least some complainants would welcome updates and may use a point of contact at the provider if they had one. To the extent that further information may be needed to consider a complaint, this would facilitate its provision.

16.105 Singapore's Code of Practice for Online Safety requires that where a large social media service receives a report that is not frivolous or vexatious, the end-user who submitted the report must be informed of the service provider's decision and action taken with respect to that report without undue delay.<sup>379</sup>

16.106 The Australian Social Media Online Safety Code requires at a minimum that *"a provider of a Tier 1 or Tier 2 social media service [excluding low risk services] must ensure that an Australian end-user who makes a report or complaint is informed in a reasonably timely manner of the outcome of the report or the complaint."*<sup>380</sup>

16.107 Recommending that services engage with users on their complaints and/or telling complainants the outcome of complaints would reassure users that they were considered, and would be likely to encourage future complaints. Providing outcomes may help to educate users on what content was and was not violative, which over time may help to improve the quality of complaints.

16.108 However, the recommendation may go further than is required to achieve this. Nor do we have sufficient evidence at this stage of the practicalities and costs of implementing such a requirement at scale for each of the types of complaints that services are required to consider. While we welcome further evidence on the topic to inform our future work, at this stage we are not proposing that transparency and accountability would require services to do this.

---

<sup>379</sup> Infocomm Media Development Authority, 2023. [Singapore's Online Safety Code of Practice](#).

<sup>380</sup> Australian E-Safety Commissioner 2023. [Australian Social Media Online Safety Code p17](#).

## Costs and risks

### Acknowledging complaints and providing timeframes

- 16.109 Services would incur costs relating to informing the complainant that the complaint has been received and with an indicative timeframe for handling the complaint. However, we consider these are likely to be small.
- a) We expect that any services that receive more than a small number of complaints would want to automate this response (e.g. through an email or pop-up message). We have estimated that the direct costs of this measure would take approximately 5 to 50 days of software engineering time, with potentially up to the same again in non-engineering time. Using our assumptions on labour costs required for this type of work set out in Annex 14, we would expect the one-off direct costs to be somewhere in the region of £2,000 to £50,000. There would also be some ongoing costs involved in maintaining this. We expect that services with less complex systems and governance processes are likely to incur costs at the lower end of this range, which is likely to be the case for smaller services.
  - b) Small and low risk services that do not receive any or very few complaints may choose to have a manual approach to sending acknowledgements and indicative timescales.
- 16.110 The indicative timeframes would not be binding on the service. However, there may be some indirect impacts from providing indicative timeframes. For example, a service may receive more repeat complaints if the indicative timeframe were not met. Its users may also be more dissatisfied with the service. This would incentivise the service concerned to meet its own indicative timeframes but may also incentivise it to set longer ones than it otherwise might.
- 16.111 While there may be some benefits to complainants in providing a point of contact or a way to check the status of complaints, we have little information on the costs of doing this and consider that they could be very significant given the volume of complaints the largest services receive. It is not clear the overall impact on user safety is sufficient to make this proportionate.

### Rights impacts

#### *Impacts on freedom of expression and applicable safeguards*

- 16.112 Our proposals in this section are only about services' communications with users. We do not consider our proposed recommendations would have any impact on users', affected persons' rights to freedom of expression. To the extent that our recommendations ask services to convey information they might not otherwise convey, there is a potential small impact on services' rights to freedom of expression. However, we consider this proportionate in the interests of protecting the rights of users and affected persons, in particular their rights to have their complaints handled appropriately and in the light of the Act's objectives on transparency and accountability.
- 16.113 To the extent that a service needed to retain information to process complaints, this may include personal data. However, we are not proposing to recommend that services should process or retain any extra information beyond the minimum needed to comply with duties which are set out clearly on the face of the Act. To the extent that services choose to do so, this data would be held by the service subject to data protection laws.

## Provisional conclusion

- 16.114 Sending an acknowledgement with an indicative timeframe for considering complaints will signal to complainants that their complaints are being dealt with. This should reassure people and encourage them to make complaints in the future. Where complaints are about illegal content, this should mean more illegal content is complained about by users and identified on the service (or in search results). For this reason, we consider there are likely to be benefits from this proposed measure. These benefits will tend to be greater the more complaints a service receives.
- 16.115 For services that automate the sending of these messages, there will be a direct cost that is largely one off. As the indicative timescale would not be binding, services would retain flexibility in how they prioritised different complaints.
- 16.116 On balance, we consider that this measure is likely to be proportionate for all services. Our analysis suggests the costs would be relatively small. In light of the evidence that an absence of clarity about timelines and process for addressing complaints deters complainants (consequently reducing detection of illegal content), it appears that the benefits of applying this measure to large and risky services could be relatively significant. For services that receive very few complaints, the benefits would be small, but the costs would likely also be low as such services could retain a manual process for acknowledging complaints and sending an indicative timeframe.
- 16.117 We therefore propose recommending that our Codes on CSEA, terror, and other duties should say that all services should acknowledge receipt of a relevant complaint and provide the complainant with an indicative timeframe for deciding the complaint. See Recommendation 5C in our draft Codes in Annexes 7 and 8.

## Measure 4: appropriate action in response to complaints made on user-to-user services

---

### Harms that the measure seeks to address

- 16.118 The Act requires all regulated U2U services to operate processes that provide for appropriate action to be taken in response to reports about illegal content and other types of complaints. The appropriate action that a service might take will depend on the type of complaint.
- 16.119 We have therefore considered what ‘appropriate action’ might mean for U2U services in the context of the different types of complaints envisaged by the Act:
- a) Illegal content complaints
  - b) Wrongful takedown/blocking
  - c) Non-compliance with safety duty or content reporting duty
  - d) Non-compliance with freedom of expression or privacy duty
  - e) The use of proactive technology to moderate content
- 16.120 It is important to note key interdependencies and links between the operation of the recommended measures below to other recommendations set out in other chapters to this consultation, as well as Ofcom guidance, notably:

- a) Governance and accountability
- b) Content moderation
- c) Terms of Service (within Terms of Service and Publicly Available Statements)
- d) Ofcom's Illegal Content Judgements Guidance

16.121 The proposals set out below are Ofcom's recommended steps that a service provider should take in response to complaints. Because the measures are a package, we consider the costs and rights implications of our proposals after we have set out our thinking in relation to each type of complaint.

### Appropriate action in response to complaints about illegal content

16.122 The Act creates complaints handling duties in relation to UK users and affected persons. This does not, of course, prevent a service from offering a complaints handling process for all of its users in other jurisdictions, and we expect that many will continue to do so for commercial reasons. However, Ofcom's Codes will not be relevant for non-UK complaints.

16.123 If a service wishes to limit the complaints it considers to those it is required by the Act to consider it would first need to know if the user who has submitted an illegal content complaint has been served this content in the UK. As discussed above at para 16.22 one way of achieving this might be for a service to provide, as part of its complaints process, a way for users to indicate from which jurisdiction they are filing their complaint.

16.124 Once a complaint has been received, it should enter the service's content moderation function. As set out in Chapter 12 – U2U Content Moderation, this means that all services will need to handle the complaint as suspected illegal content under Measure 1. If they are satisfied that their terms and conditions secure that all illegal content is prohibited, they can apply their terms and conditions. If not, they need to make an illegal content judgment. Large services and smaller services that have significant risks would also need to handle the complaint in accordance with their prioritization process and performance targets under U2U Content Moderation Measures 3 and 4.

16.125 Smaller services which are low risk for illegal content may not receive many, if any, complaints. But the number of complaints such services receive could vary greatly depending on their business models and user base. Some small and low risk services may still need or want to establish a prioritisation process and associated performance targets for their content moderation function, in order to manage their workflow – this may be the least onerous and most effective approach even for a tiny service, if it predictably receives a large volume of complaints. However, for a service which receives hardly any complaints, it may be less burdensome and equally effective for it simply to process promptly every complaint it receives. Therefore, we consider that if a small and low risk service has elected to establish a prioritization process and performance targets for itself, it would be appropriate to abide by them. But a service which has none would need to process all complaints received promptly.

16.126 In either case, Measure 1 in Chapter 12 – U2U Content Moderation would then apply in relation to any content moderation decisions about the content. In other words, the service would either need to make an illegal content judgment or, if it is satisfied that its terms and conditions prohibit the types of illegal content defined in the Act which are relevant to the complaint, consider whether the content is in breach of those terms of service. If the

content was either illegal content or in breach of the relevant term of service, it would need to be taken down swiftly for UK users.

16.127 For the reasons set out above, we therefore propose to recommend that when a service receives a relevant complaint about suspected illegal content, then:

- a) if the service has established a process for content prioritisation and applicable performance targets, it should handle the complaint in accordance with them; or
- b) if the service has no process for content prioritisation and applicable performance targets it should consider the complaint promptly; and
- c) in either case, it should comply with Measure 1 in the content moderation duties regarding takedown.

### Appropriate action in response to complaints about the wrongful takedown of content on the basis it is illegal content, and wrongful user restrictions (including blocking of user access to a service) on the basis of content being illegal content

16.128 Services have a duty to handle complaints from UK users when their content has been taken down on the basis that it is illegal content, and also complaints where a UK user has been blocked or restricted because a service believes that content being shared by them is illegal content. In other words, where a user's content has been taken down on the basis that the service has judged it to be illegal content and the user considers this judgement to be erroneous, the service must offer the user the opportunity to complain and must handle their complaint. We refer to these types of complaints as 'appeals'. We consider appeals to be a means of protecting users against excessive takedown of content, and interpret what is 'appropriate action' in response to these complaints in the light of the importance of users' rights to freedom of expression.

16.129 Until other provisions of the Act are brought into force<sup>381</sup>, a service is only required by the Act to handle appeals when it has made an illegal content judgement.<sup>382</sup> It may however choose to handle all, or a wider range of, content takedown or user restriction complaints.

16.130 A service should consider appeals where the original decision was made on the basis that content was illegal content. Therefore, if it does not wish to handle all other complaints about takedown and action against users, this means it needs to find a way to identify those complaints. If services notify users that action has been taken because the service has judged content to be illegal content, we consider it likely to be a straightforward matter to build this into complaints forms. If they choose instead to retain records of their decisions, they will need to be able to match the decisions to the complaints received.

16.131 Some services may choose to run appeals through their main content moderation function. Others may establish a separate team. In either case, questions arise for services about how

---

<sup>381</sup> Section 21 contains further complaints handling duties for services likely to be accessed by children. We intend to consult on Codes for this in due course. Sections 71 and 72 of the Online Safety Act will create duties for U2U services to deal with complaints that are not caught by the duty in section 21. Ofcom does not have a Code-making function in relation to these.

<sup>382</sup> This means that, in principle, where a service has taken down a piece of content because it violated the service's terms of service, rather than because it has judged the content to be illegal, the Act does not require the service to allow the user that posted the content to appeal. We recognise, however, that many services will choose to offer users the ability to appeal in such circumstances and would encourage them to do so.

quickly it is appropriate to review the decision, and what priority to give it as against other decisions.

- 16.132 For services that are low risk and are not large, our provisional view for the reasons set out in paragraph 12.143 of the U2U content moderation chapter regarding proportionality, is that there is no need to make detailed recommendations in Codes on prioritisation.
- 16.133 For large services and for services that are multi-risk, however, we consider the volumes of content they are likely to need to consider are such that users may be harmed if they do not consider appropriate prioritisation in advance. We provisionally consider that large services and services that are multi-risk should have regard to the following matters in determining what priority to give to review of the complaint:
- a) the severity of the action taken against the user as a result of the decision that the content was illegal content;
  - b) whether the decision that the content was illegal content was made by proactive technology and the likelihood of false positives generated by the specific proactive technology used; and
  - c) the service's past error rate in making illegal content judgments of the type concerned.
- 16.134 On the timeliness of considering appeals, for all the reasons set out in paragraph 12.143 of the U2U content moderation chapter, we do not consider it appropriate for Ofcom to make specific recommendations. For services which are low risk and not large, which we expect will not receive many complaints, let alone many appeals, we consider it will be sufficient to say that appeals should be determined promptly.
- 16.135 However, we consider that taking this approach for large services and services that are multi-risk could create perverse incentives and lead to user harm. We therefore propose to recommend that such services should include in their content policies, targets as to speed and accuracy for the determination of appeals. Similar recommendations in Chapter 12 as to monitoring and resourcing would apply in relation to these too, for the reasoning given there.
- 16.136 We consider that if, on review, a service reverses a decision that content was illegal content, in principle the service should:
- a) restore the content and/ or the user's account to the position they would have been in had the content not been judged to be illegal content; and
  - b) where necessary to avoid similar errors in future, adjust the relevant content moderation guidance.
- 16.137 There is a risk that automated content moderation technology may be involved in a takedown or downranking decision. We therefore propose that if on review, a service reverses a decision that content was illegal content, then where necessary to avoid similar errors in future, the service should take such steps as are within its power to secure that the use of automated content moderation technology does not cause the same content to be taken down again.

### Appropriate action in response to complaints relating to the use of proactive technology

- 16.138 The Act requires services to take appropriate action in response to complaints about the use of proactive technology on that service when:



- a) the use of proactive technology on the service results in content being taken down or access to it being restricted, or given a lower priority or otherwise becoming less likely to be encountered by other users; and
- b) the user considers that the proactive technology has been used in a way not contemplated by, or in breach of, the terms of service (for example, by blocking content not of a kind specified in the terms of service as a kind of content in relation to which the technology would operate).

16.139 This category of complaint is particularly broad, because it applies to all kinds of proactive technology and could affect any type of content, not just illegal content.

- a) We consider that complaints about wrongful takedown of content on the basis that it is illegal content should be handled in accordance with paragraphs 16.128-16.137 above regardless of whether the takedown decision concerned was made by a human or by technology. If services notify users when proactive technology has been used, a user will know when an illegal content judgement has been made and will be in a position to make an appropriate complaint.
- b) If the service does not notify but retains records, it will be able to identify the complaint accordingly.
- c) If the service has no information on whether or not proactive technology was used in relation to the content, it should assume that it was, and handle the complaint accordingly.

16.140 But we also need to think about complaints about proactive tech being used inconsistently with terms and conditions, where there has *not* been a decision that content is illegal content. For those, we are of the view that the reference in the Act to terms and conditions in the definition of this complaint type<sup>383</sup> makes it clear that the proper basis of a complaint about the use of proactive technology is not necessarily about the nature of the content taken down or the fact of the proactive technology having been used, but whether the operation of the proactive technology concerned is consistent with the terms of service. This reflects how proactive technology is addressed in the safety duty (section 10 (7)), and how terms and conditions are addressed in the remainder of the Act (section 71 of the Act will place duties on Category 1 user-to-user services not to act against users except in accordance with terms of service.)

- a) If services notify users when proactive technology has been used, a user will know when an illegal content judgement has been made and will be in a position to make an appropriate complaint.
- b) If the service does not notify but retains records, it will be able to identify the complaint accordingly.
- c) If the service has no information on whether or not proactive technology was used in relation to the content, it should assume that it was, and handle the complaint accordingly.

16.141 Where a complaint is made about a content moderation decision made through the use of proactive technology on a service which is not a Category 1 service (as will be defined by the Secretary of State in due course), and where that complaint does not relate to an illegal

---

<sup>383</sup> 21(4)(e) of the Act

content judgment (but the technology was potentially used outside the parameters set out in published terms of service), we provisionally consider that the appropriate action by the provider concerned would be no more than to inform the user of their right, if they consider that the service is in breach of contract, to bring a claim for breach of contract.<sup>384</sup> For now, there are no Category 1 services, so this would account for all complaints.

## Appropriate action in response to complaints about compliance with illegal content duties, illegal content reporting, freedom of expression or privacy

- 16.142 At this stage, we do not consider that we are in a position to predict with sufficient certainty the many different types of complaint that may be submitted to services in relation to compliance with the safety duties, the reporting duty, freedom of expression or privacy, or to set out what action is appropriate in relation to each of them. Consequently, we are not currently proposing to make detailed recommendations in Codes as to what final action may be appropriate in relation to the handling of most of these complaint types, although we will keep this position under review.
- 16.143 We also note that there is a significant risk of overlap between complaints about compliance with the safety duties and freedom of expression, and complaints about illegal content, wrongful takedown or blocking, or use of proactive technology inconsistently with terms and conditions. Where a complaint falls into one of those categories as well as this, we provisionally consider it appropriate for the service to handle it in accordance with our proposed recommendations for those complaint types.
- 16.144 However, we do not think we need to specify this in a specific measure because we provisionally think the appropriate action for services in relation to complaints concerning compliance with illegal content duties, illegal content reporting, freedom of expression and privacy would be to establish a triage process with a view to protecting users from harm, including harm to their rights. A responsible person, team or function for such complaints should be nominated to lead this triage process and ensure complaints reach the most relevant function or team. They should be dealt with in a way that protects users and the service's compliance with other applicable laws in question, within timeframes the service has determined are appropriate, and in accordance with our other proposed Code measures relating to complaints.

## Costs and risks

- 16.145 The costs of taking appropriate action for complaints will vary across different types and sizes of services, and for services with different levels of risk. While we expect the costs to be very significant for some service providers, these costs are mitigated by us not proposing to set specific timescales for looking at complaints in our proposals for appropriate action for complaints and content moderation.
- 16.146 Also, while recognising it depends on the nature of the service, we would generally expect the potential volume of complaints about illegal content to vary with the size of the service.

---

<sup>384</sup> Section 72(1) of the Act provides that U2U services must include clear and accessible provisions in the terms of service informing users about their right to bring a claim for breach of contract if—(a) regulated user-generated content which they generate, upload or share is taken down, or access to it is restricted, in breach of the terms of service, or (b) they are suspended or banned from using the service in breach of the terms of service. Ofcom has no Code-making duty or power in relation to this provision.

Complaints about illegal content are likely to vary with the volume of content being shared by users. For most services, we anticipate the largest volume of complaints caught by the complaints handling duty to be complaints about illegal content, and these costs could be regarded as part of content moderation. Complaints about content moderation decisions will tend to vary with the total volume of content moderation decisions. If costs tend to vary with the size of the services, it means services with the highest costs will tend to be those with the greatest ability to bear those costs.

- 16.147 While the costs may be significant for some services, we believe they are imposed, in large part, by aspects of the Act in relation to which in practice Ofcom has little discretion. Of the complaints types covered by this duty, the majority are likely to relate to suspected illegal content, and a service would not be able to comply with the takedown duty in section 10(3)(b) of the Act if it did not consider them.
- 16.148 Additionally, we have considered the potential added complexity for all kinds of services in making judgements about illegal content. However, as set out above the Act does not necessarily require services to make illegal content judgments if they are satisfied that their terms of service or community guidelines prohibit content that would be considered illegal in the UK. To the extent that new illegal content judgments are required, this is down to the requirements of the Act.
- 16.149 The duty to consider appeals is a key way in which the Act safeguards users' rights to freedom of expression so we consider that our discretion in determining what could be said to be 'appropriate' for appeals is not wide. For other types of complaints, we are proposing what we see as the minimum requirements which could be consistent with the duty as set out in the Act.
- 16.150 Due to the fact that the reporting and complaints duties apply to all in-scope services, we have proposed setting out broad features (as opposed to specific ones) that we recommend services consider when designing their reporting and complaints systems, and believe we have approached this in a way that seeks as far as possible to elucidate (and not build on) the basic legal requirements set out in the Act. On this basis, we believe our proposals are proportionate and suitable for a very wide range of services.

## **Rights impacts**

- 16.151 We do not consider our proposed recommendations would have any negative impact on users', affected persons' or services' rights to freedom of expression.
- 16.152 To the extent that the complaints handling duty relates to appeals against wrongful takedown or restriction of users, we see them as an important safeguard of users' rights.
- 16.153 To the extent that a service needed to retain information to process complaints, this may include personal data. However, we are not proposing to recommend that services should process or retain any extra information beyond the minimum needed to comply with duties which are set out clearly on the face of the Act. To the extent that services choose to do so, this data would be held by the service subject to data protection laws.

## **Provisional conclusion**

- 16.154 As set out above, the appropriate handling of complaints by U2U services is required by section 21 of the Act, and so this proposed measure would apply to all U2U services.

16.155 Section 21 is not part of the safety duty, and so this proposed measure belongs in our Code on other duties. For all the reasons above, we therefore propose to recommend that this Code contains the provisions set out as Recommendations 5D-5H in our draft Code in Annex 7.

16.156 However, we consider that handling complaints about illegal content, including where they are received through a user report, is also necessary for a service to meet its safety duties in relation to CSEA and terrorism content<sup>385</sup> and we therefore also propose to also include this measure in our CSEA and terrorism Codes.

## Measure 5: appropriate action in response to complaints made on search services

---

### Harms that the measure seeks to address

16.157 Section 32(4) of the Act requires all regulated search services to operate processes that provide for appropriate action to be taken in response to complaints. The appropriate action that a service might take will depend on the type of complaint.

16.158 We have therefore considered what ‘appropriate action’ might mean for search services in the context of the different types of complaints envisaged by the Act:

- a) Illegal content complaints
- b) Wrongful deindexing/downranking
- c) Non-compliance with safety duty
- d) Non-compliance with freedom of expression or privacy duty
- e) The use of proactive technology to moderate content.

16.159 It is important to note key interdependencies and links between the operation of the recommended measures below to other recommendations set out in other chapters to this consultation, as well as Ofcom guidance, notably:

- a) Governance and accountability
- b) Publicly Available Statements (within Terms of Service and Publicly Available Statements)
- c) Search moderation
- d) Search service design
- e) Ofcom’s Illegal Content Judgements Guidance

16.160 The proposals set out below are Ofcom’s recommended steps that a search service provider should take in response to reports about illegal content and other relevant complaints.

---

<sup>385</sup> Specifically, for U2U services, their duties relating to minimising the length of time for which any priority illegal content is present; and where the provider is alerted by a person to the presence of any illegal content, swiftly take down such content (section 10(3)). For search services, their duties relating to minimising the risk of individuals encountering search content which is priority illegal content (section 27(3)).

## Appropriate action in response to complaints made about illegal content appearing in search results

- 16.161 The Act creates complaints handling duties in relation to UK users and affected persons. This does not, of course, prevent a service from offering a complaints handling process for all of its users in other jurisdictions, and we expect that many will continue to do so for commercial reasons. However, Ofcom's Codes will not be relevant for non-UK complaints.
- 16.162 If a service wishes to limit the complaints it considers to those it is required by the Act to consider it would first need to know if the user who has submitted an illegal content complaint has been served this content in the UK. As discussed above at paragraph 16.26 one way of achieving this might be for a service to provide, as part of its complaints process, a way for users to indicate from which jurisdiction they are filing their complaint.
- 16.163 Once a complaint has been received, it should enter the service's search moderation function. As set out in chapter 13, this means that large (or multi-risk) general search services will need to prioritise the complaint in accordance with their prioritisation process and performance targets.
- 16.164 Smaller services which are low risk for illegal content, which may include vertical search services, may not receive many, if any, complaints. But the number of complaints such services receive could vary greatly depending on their business models and user base. Some small and low risk services may still need or want to establish a prioritisation process and associated performance targets for their search moderation function, in order to manage their workflow – this may be the least onerous and most effective approach even for a microbusiness if it predictably receives a large volume of complaints.
- 16.165 However, for a service which receives a very low number of complaints (if any), it may be less burdensome and equally effective for it simply to process promptly every complaint it receives. Therefore, we consider that if a small and low risk service has elected to establish a prioritisation process and performance targets for itself, it would be appropriate to work towards achieving them. But a service which has none would need to process all complaints received promptly.
- 16.166 In either case, Measure 1 in Chapter 13 on search moderation would then apply in relation to any search moderation decision about the URL or search results in question. In other words, the service would either need to make an illegal content judgment or, if it was satisfied that its publicly available statement prohibited the types of illegal content defined in the Act which were relevant to the complaint, consider whether the content is in breach of the publicly available statement. If the content was either illegal content or in breach of the publicly available statement, it would need to be deindexed or downranked for UK users.
- 16.167 For the reasons set out above, we therefore propose to recommend that when a complaint about suspected illegal search content is submitted by a user or interested person, then:
- a) if the service has established a process for search moderation prioritisation and applicable performance targets, it should handle the complaint in accordance with them; or
  - b) where a service has no process for prioritisation and applicable performance targets it should consider the complaint promptly; and
  - c) in either case, it should comply with Measure 1 in chapter 13 on search moderation regarding deindexing or downranking illegal content.

16.168 As explained in paragraph 11.65, downstream search services are general search services that do not produce their own index or ranking of search content that might be accessed via their search engine. It would therefore not be possible for them to deindex or downrank content in response to an illegal content complaint. However, we consider that this measure should apply to them similarly, since they can secure by contract that complaints are dealt with appropriately.

### Appropriate action in response to complaints about suspected deindexing or downranking of content because it is thought to be illegal content

16.169 Services have a duty to handle complaints from interested persons when search content has been deindexed or downranked on the basis that it is considered to be illegal content in the UK. For ease of reading, and as set out above, we have been using the general term ‘website owner’ in this chapter to refer to interested persons. For the purposes of considering this type of complaint, it is useful to remember its full definition: ‘interested person’ means 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.

16.170 Until other provisions of the Act are brought into force<sup>386</sup>, a service is only required by the Act to handle these complaints when it has made an illegal content judgement. It may however choose to handle all or a wider range of complaints.

16.171 If the service does not wish to handle all complaints about takedown and action against interested persons, it needs to find a way to identify complaints where the decision was made on the basis that content was illegal content. As noted above, search services will not usually be in a position to identify or communicate with the provider of a URL or database, so they are unlikely to be in a position to notify interested persons of this. However, we consider that services are likely to know when they have made a decision that search content is illegal content for the purposes of the Act.

16.172 Some search services may choose to run appeals through their main search moderation function. Others may establish a separate team. In either case, questions arise for services about how quickly it is appropriate to review the decision, and what priority to give it as against other decisions.

16.173 For services that are low risk and are not large, our provisional view for the reasons set out in paragraph 13.121 of the search moderation chapter is that there is no need to make detailed recommendations in Codes on prioritisation.

16.174 For large services (apart from vertical search services) and for services that are multi-risk, however, we consider the volumes of content they are likely to need to consider are such that interested persons may be harmed if they do not consider appropriate prioritisation in advance. We provisionally consider that large services that are not vertical search services and services that are multi-risk should have regard to the following matters in determining what priority to give to review of the appeal:

---

<sup>386</sup> Section 32 contains further duties for services likely to be accessed by children. We intend to consult on Codes for this in due course.

- a) the severity of the action taken against the interested person as a result of the decision that the content was illegal content;
  - b) whether the decision that the content was illegal content was made by proactive technology and the likelihood of false positives generated by the specific proactive technology used; and
  - c) the service's past error rate in making illegal content judgments of the type concerned.
- 16.175 On the timeliness of considering appeals, for all the reasons set out in paragraph 13.66, we do not consider it appropriate for Ofcom to make specific recommendations. For services which are low risk and not large, which we expect will not receive many complaints, let alone many appeals, we consider it will be sufficient to say that appeals should be determined promptly.
- 16.176 However, we consider that taking this approach for large services (other than vertical search services) and services that are multi-risk could create perverse incentives and lead to harm. We therefore propose to recommend that such services should include in their content policies, targets as to speed and accuracy for the determination of appeals. Our recommendations in Chapter 13 as to monitoring and resourcing would apply in relation to these too, for the reasoning given there.
- 16.177 We consider that a search service should have regard to a number of factors in determining what priority to give to review of the complaint, and we have set these out below.
- a) the severity of the action taken against the interested person as a result of the decision that the content was illegal content;
  - b) whether the decision that the content was illegal content was made by proactive technology; and
  - c) the service's past error rate in making illegal content judgments of the type concerned.
- 16.178 A service should review the illegal content judgement it made, having regard to any new information it holds.
- 16.179 If, on review, a service reverses a decision that a URL or database contained illegal content, in principle the service should:
- a) restore the content to the position it would have been in had the content not been judged to be illegal content; and
  - b) where necessary to avoid similar errors in future, adjust the relevant moderation guidance; and
  - c) where necessary to avoid similar errors in future, take such steps as are within its power to secure that the use of automated moderation technology does not cause the same content to be deindexed or deprioritised again.

### Appropriate action in response to complaints by an interested person about suspected deindexing or downranking of URLs due to the use of proactive technology

- 16.180 Search services must take appropriate action in response to complaints by an 'interested person' if:



- a) the use of proactive technology on that search service results in content relating to that interested person being deindexed, or downranked; and
  - b) the interested person believes that proactive technology has been used in a way not contemplated by, or in breach of, the search provider’s policies on its use (for example, by affecting content not of a kind specified in those policies being subject to the technology’s operation)
- 16.181 This category of complaint is particularly broad, because it applies to all kinds of proactive technology and could affect any type of content related decision, not just those relating to illegal content.
- 16.182 We consider that complaints about wrongful deindexing or downranking of a URL on the basis of a service identifying illegal content should be handled in accordance with paragraphs 16.169-16.179 above regardless of whether the takedown decision concerned was made by a human or by technology.
- 16.183 But we also need to think about complaints about proactive tech being used inconsistently with terms and conditions, where there has not been a decision that content is illegal content. For those, we are of the view that the reference in the Act to terms and conditions in the definition of this complaint type<sup>387</sup> makes it clear that the proper basis of a complaint about the use of proactive technology is not necessarily about the nature of the content in question, but whether the operation of the proactive technology concerned is consistent with the terms of service. This reflects how proactive technology is addressed in the safety duty (section 10(7)), and how terms and conditions are addressed in the remainder of the Act (section 71 of the Act will place duties on Category 1 user-to-user services not to act against users except in accordance with terms of service.)
- a) If services notify interested persons when proactive technology has been used, the interested person will know when an illegal content judgement has been made and will be in a position to make an appropriate complaint.
  - b) If the service does not notify interested persons, but instead retains records, it will be able to identify the complaint accordingly.
  - c) If the service has no information on whether or not proactive technology was used in relation to the content, it should assume that it was, and handle the complaint accordingly.
- 16.184 Search services cannot be Category 1 services. Where a complaint is made about a deindexing or downranking decision made through the use of proactive technology on a search service, and where that complaint does not relate to an illegal content judgment (but the technology was potentially to have been used outside the parameters set out in published terms of service), we provisionally consider that the appropriate action by the provider concerned would be no more than to inform the interested person of their rights – for example, if they consider that the service is in breach of contract, they could bring a claim for breach of contract.

---

<sup>387</sup> 21(4)(e) of the Act.

## Appropriate action in response to complaints about compliance with illegal content duties, illegal content reporting, freedom of expression or privacy

- 16.185 At this stage, we do not consider that we are in a position to predict with sufficient certainty the many different types of complaint that may be submitted to services in relation to compliance with the safety duties, the reporting duty, freedom of expression or privacy, or to set out what action is appropriate in relation to each of them. Consequently, we are not currently proposing to make detailed recommendations in Codes as to what final action may be appropriate in relation to the handling of most of these complaint types, although we will keep this position under review.
- 16.186 We also note that there is a significant risk of overlap between complaints about compliance with the safety duties and freedom of expression, and complaints about illegal content, wrongful deindexing or downranking, or use of proactive technology inconsistently with terms and conditions. Where a complaint falls into one of those categories as well as this, we provisionally consider it appropriate for the service to handle it in accordance with our proposed recommendations for those complaint types.
- 16.187 However, we do not think we need to specify this in a specific measure because we provisionally think the appropriate action for services in relation to complaints concerning compliance with illegal content duties, illegal content reporting, freedom of expression and privacy would be to establish a triage process with a view to protecting users and interested persons from harm, including harm to their rights. A responsible person, team or function for such complaints should be nominated to lead this triage process and ensure complaints reach the most relevant function or team. They should be dealt with in a way that protects users and the service's compliance with other applicable laws in question, within timeframes the service has determined are appropriate, and in accordance with our other proposed Code measures relating to complaints.

## Costs and risks

- 16.188 The costs of taking appropriate action for complaints will vary across different types and sizes of services, and for services with different levels of risk. While we expect the costs to be very significant for some service providers, these costs are mitigated by us not proposing to set specific timescales for looking at complaints in our proposals for appropriate action for complaints and search moderation.
- 16.189 Also, while recognising it depends on the nature of the service, we would generally expect the potential volume of complaints about illegal content to vary with the size of the service. Complaints about illegal content are likely to vary with the volume of search queries users run. For most services, we anticipate the largest volume of complaints caught by the complaints handling duty to be complaints about illegal content, and these costs could be regarded as part of search moderation. If costs tend to vary with the size of the services, it means services with the highest costs will tend to be those with the greatest ability to bear those costs.
- 16.190 While the costs may be significant for some services, we believe they are imposed, in large part, by aspects of the Act in relation to which in practice Ofcom has little discretion. Of the complaints types covered by this duty, the majority are likely to relate to suspected illegal content, and a service would not be able to comply with the duty in section 27(3) of the Act if it did not consider them.

- 16.191 Additionally, we have considered the potential added complexity for all kinds of services in making judgements about illegal content. However, as set out above the Act does not necessarily require services to make illegal content judgments if they are satisfied that their terms of service prohibit content that would be considered illegal in the UK. To the extent that new illegal content judgments are required, this is down to the requirements of the Act.
- 16.192 The duty to consider appeals is a key way in which the Act safeguards users' rights to freedom of expression so we consider that our discretion in determining what could be said to be 'appropriate' for appeals is not wide. For other types of complaints, we are proposing what we see as the minimum requirements which could be consistent with the duty as set out in the Act.
- 16.193 Due to the fact that the reporting and complaints duties apply to all in-scope services, we have proposed setting out broad features (as opposed to specific ones) that we recommend services consider when designing their reporting and complaints systems, and believe we have approached this in a way that seeks as far as possible to elucidate (and not build on) the basic legal requirements set out in the Act. On this basis, we believe our proposals are proportionate and suitable for a very wide range of services.

## Rights impacts

- 16.194 We do not consider our proposed recommendations would have any negative impact on users', affected persons', interested persons' or services' rights to freedom of expression.
- 16.195 To the extent that the complaints handling duty relates to appeals against wrongful deindexing or deprioritisation of search content, we see them as an important safeguard of interested persons' rights.
- 16.196 To the extent that a service needed to retain information to process complaints, this may include personal data. However, we are not proposing to recommend that services should process or retain any extra information beyond the minimum needed to comply with duties which are set out clearly on the face of the Act. To the extent that services choose to do so, this data would be held by the service subject to data protection laws.

## Provisional conclusion

- 16.197 As set out above, the appropriate handling of complaints by search services is required by section 32 of the Act, and so this proposed measure would apply to all search services.
- 16.198 Section 32 of the Act is not part of the safety duty, and so this proposed measure belongs in our Code on other duties. For all the reasons above, we therefore propose to recommend that this Code contains the provisions set out as Recommendations 5D to 5H in the draft Code.
- 16.199 However, we consider that handling complaints about illegal content, including where they are received through a user report, is also necessary for a service to meet its safety duties in relation to CSEA and terrorism content<sup>388</sup> and we therefore also propose to also include this measure in our CSEA and terrorism Codes.

---

<sup>388</sup> Specifically, for U2U services, their duties relating to minimising the length of time for which any priority illegal content is present; and where the provider is alerted by a person to the presence of any illegal content, swiftly take down such content (section 10(3)). For search services, their duties relating to minimising the risk of individuals encountering search content which is priority illegal content (section 27(3)).

## Measure 6: Dedicated Reporting Channels for services with risks of fraud – U2U and search services

---

### Harms that the measure seeks to address

- 16.200 Section 10(3)(a) of the Act creates a duty to operate a U2U service using proportionate systems and processes designed to minimise the length of time for which any priority illegal content is present, and section 23(3) of the Act creates a duty to operate a search service using proportionate systems and processes designed to minimise the risk of individuals encountering search content that is illegal content.
- 16.201 Ofcom has explored the potential benefits of recommending the use of Dedicated Reporting Channels ('DRCs') in the context of these provisions.
- 16.202 The users of DRCs are often referred to as reporters or trusted flaggers. Trusted flaggers are typically entities, and not individual users, that have particular expertise and competence for the purposes of detecting, identifying and notifying services about illegal content. Trusted flaggers represent collective interests (typically through a public mandate) and normally operate independently from any online service.
- 16.203 We note that DRCs are often used by industry for a number of different types of harms. Indeed, the EU's Digital Services Act imposes a requirement on providers of online platforms that are not micro or small enterprises as defined in that legislation (unless they are very large Online Platforms) to act on information supplied by trusted flaggers 'with priority and without delay'.<sup>389</sup>
- 16.204 In theory, DRCs may be a useful means of tackling many types of harm. However, rather than focusing on DRCs in general, we have in the first instance chosen to focus our work specifically on the possibility of recommending the creation of a DRC related to fraud. This is because, following stakeholder engagement, a number of expert organisations noted various challenges with reporting fraud into online services. We therefore considered that to begin with, developing a DRC for fraud was likely to represent the most effective and proportionate response to those concerns with a view to optimising the reporting environment for online fraud. It is also important to ensure that the entities which would be entitled to use a DRC have appropriate intelligence and expertise for it to be valuable, and that they would use it responsibly. Recognising the availability of a distinct list of expert trusted flaggers with the skills and knowledge to recognise fraudulent content, we see a particular opportunity to achieve significant improvements in user safety in the light of our priority to work with other agencies in support of efforts to tackle online fraud.
- 16.205 We believe that we have sufficient evidence to allow us to develop recommendations for the UK in respect of DRCs in the context of reporting fraud, and we set out our thinking on this below.
- 16.206 According to the National Economic Crime Centre, fraud, both online and offline, is the most frequently experienced crime in the UK.<sup>390</sup> Fraud currently accounts for over 40% of all crime in the UK and this figure is growing year on year.<sup>391</sup> Action Fraud reported £2.35bn in

---

<sup>389</sup> Articles 16 and 19 of the Digital Services Act.

<sup>390</sup> National Crime Agency, [Improving the UK's response to economic crime](#), [accessed 26 September 2023].

<sup>391</sup> Office for National Statistics, 2022. [Nature of fraud and computer misuse in England and Wales - Office for National Statistics \(ons.gov.uk\)](#), [accessed 26 September 2023].

fraud related losses in 2021-22, whilst noting that 80% of reported fraud is cyber-enabled, and that social media and encrypted messaging services as an enabler is increasing throughout all aspects of fraud.<sup>392</sup> The UK Government Fraud Strategy estimates that the total economic and social cost of fraud to individuals is £6.8 billion (2019/20).<sup>393</sup> Criminals appear to be making use of the large user reach provided by online services.<sup>394</sup> They do this to expose the public to fraudulent content, with the intent of profiting.<sup>395</sup> Research carried out by Yonder on behalf of Ofcom found that nearly 9 out of 10 adult internet users (87%) have encountered content online which they believed to be a scam or fraud.<sup>396</sup> The scale of this threat is immense.

- 16.207 Fraud is a volume crime. Online services with a large user base are particularly attractive<sup>397</sup> to criminals as they make it easy for them to reach large numbers of people at low cost<sup>398</sup> and with minimal effort.
- 16.208 However, we see several of the fraud offences as being particularly difficult for services to identify accurately without reliable contextual information. We expect the most relevant priority offence within scope is likely to be fraud by false representation (s.2 of the Fraud Act 2006) given that criminals are likely to make use of the large audience reach provided by online services to socially engineer<sup>399</sup> the public, manipulating users through the use of the likeness of a trusted brand or individual. However, the priority offences in the Act include a number of others in relation to which third party input is likely to be valuable to services. In particular, the priority offences relating to financial services are sufficiently technical and complicated that services may benefit from help from the FCA to make illegal content judgements in relation to them.
- 16.209 Much of the evidence relating to relevant fraud offences (e.g., the transfer of money, monetary instruments, and digital assets) will often not be directly observable by the in-scope service where the interaction with the victim originates or begins. In contrast, these elements are more likely to be observed by financial services providers (either via in-house transaction monitoring/intelligence gathering - or via consumer reporting), law enforcement agencies and regulators such as the FCA.
- 16.210 However, such bodies may not always be users of the service, and if they are, the usual complaints process may not enable the service to quickly verify their identity as a provider of particularly credible and important information. There is also a risk that information which

---

<sup>392</sup> Action Fraud, 2021. [Annual Assessment Fraud Crime Trends 2021-22](#), [accessed 26 September 2023].

<sup>393</sup> Home Office, 2023. [Fraud Strategy: Stopping Scams and Protecting the Public](#), page 57, [accessed 18 August 2023]

<sup>394</sup> UK Finance, 2022. [Annual Fraud Report](#), [accessed 26 September 2023].

<sup>395</sup> Justice Committee, 2022. [Justice response inadequate to meet scale of fraud epidemic - Committees - UK Parliament](#), [accessed 26 September 2023].

<sup>396</sup> Ofcom, 2023. [Scale and Impact of Online Fraud](#) [accessed 26 September 2023].

<sup>397</sup> Consumers International, 2019. [Social Media Scams: Understanding the Consumer Experience to Create a Safer Digital World](#), [accessed 26 September 2023].

<sup>398</sup> Federal Trade Commission (Fletcher, E.), 2022. [Social media a gold mine for scammers in 2021](#), [accessed 26 September 2023].

<sup>399</sup> **Definition:** Social engineering is the tactic of manipulating, influencing, or deceiving a victim in order to gain control over a computer system, or to steal personal and financial information. It uses psychological manipulation to trick users into making security mistakes or giving away sensitive information. Carnegie Mellon University, [What is Social Engineering?](#) [accessed 18 August 2023].

could prevent harm may not be prioritised appropriately if it has to be provided through the same complaints process as used by users and affected persons.

- 16.211 Additionally, the complex and fragmented nature of the internet and the wider counter-fraud ecosystem means that criminals benefit from the lack of effective and consistent engagement between industries impacted by fraud, law enforcement and other relevant stakeholders.<sup>400</sup> Our discussions with law enforcement and industry stakeholders from the banking sector have highlighted that there is a lack of consistency in how fraud investigators and expert teams are able to report fraudulent content to the largest online user-to-user services.<sup>401</sup> This means that stakeholders raising concerns about fraud are treated differently by different online services.
- 16.212 We therefore believe there is scope for services to take proportionate steps to provide a more accessible mechanism for these reports to be made. Some online services have a form of reporting process in place for public sector bodies and law enforcement. This is important for service providers in ensuring that they have the necessary infrastructure and processes in place to obtain useful information to corroborate existing suspicions or to identify instances of illegal content. As *TechUK* commented in its response to our 2020 Video-Sharing Platform Regulation Call For Evidence, “...the vast majority of platforms now operate ‘trusted flagger’ programmes. These programmes enable law enforcement, civil society, charities and other important and reliable stakeholders to alert platforms to harmful or violating content with a fast-tracked review process.”<sup>402</sup>
- 16.213 However, evidence also suggests that there is limited clarity on how trusted flaggers with relevant fraud expertise can engage effectively with online services. There are inconsistent levels of provision, with some services offering a dedicated route for fraud reporting, some offering “informal”<sup>403</sup> modes of engagement, and others offering routes for reporting that are “not suitable”.<sup>404</sup> Some engagement with online services occurs through law enforcement entities.<sup>405</sup> We also understand that, in instances where a dedicated reporting channel has not been provided, some organisations with valuable insights have resorted to using general public reporting routes for flagging suspected fraud to online services.<sup>406</sup>
- 16.214 Where a dedicated reporting facility has been implemented, the experience of trusted flaggers can be variable, due to a lack of meaningful engagement from some online services.<sup>407</sup> There is a clear call from stakeholders for *the implementation of “better relationships with platforms”*<sup>408</sup>, *with more effective communication channels.*<sup>409</sup> One stakeholder has raised the need to “reduce friction” when seeking to engage with online services for the purpose of tackling fraud.<sup>410</sup> Most notably, City of London Police has flagged that they “strongly support the creation of specific DRCs for fraud crime. This would go a long

---

<sup>400</sup> Royal United Services Institute, 2021. [The UK’s Response to Cyber Fraud](#), [accessed 31 August 2023].

<sup>401</sup> [CONFIDENTIAL X].

<sup>402</sup> [Tech UK response to 2020 Video-Sharing Platform Regulation Call For Evidence](#), p.3. [accessed 31 August 2023]

<sup>403</sup> [CONFIDENTIAL X].

<sup>404</sup> [CONFIDENTIAL X].

<sup>405</sup> [CONFIDENTIAL X].

<sup>406</sup> [CONFIDENTIAL X].

<sup>407</sup> [CONFIDENTIAL X].

<sup>408</sup> [CONFIDENTIAL X].

<sup>409</sup> [CONFIDENTIAL X].

<sup>410</sup> [CONFIDENTIAL X].



way to resolving the current difficulties faced by the force: navigating a patchwork of industry contact points; with varying levels of assistance and buy-in".<sup>411</sup> This challenge with navigating a patchwork of contacts and varying engagement is also reinforced by other feedback that we have received, noting the difficulties experienced by expert organisations when seeking to engage with larger services that are owned and operated in different jurisdictions.<sup>412</sup> One stakeholder has even observed an "unwillingness of online services to engage with key groups/forums".<sup>413</sup>

16.215 It is evident that there is inconsistency in how online services engage with different external entities or trusted flaggers, and variation in how they facilitate the reporting of fraud by these organisations. This is leading to growing frustration across various industries and missed opportunities to tackle the rising issue of fraud.

16.216 **In response to our 2022 Illegal Harms Call for Evidence, one stakeholder asked for 'effective, quick communication channels' to enable expert organisations to alert online services of suspected fraud.**<sup>414</sup> Similarly, various stakeholders in the banking sector have called for "direct access to moderators for regulated sectors".<sup>415</sup> Although the banking sector has called for direct routes into content moderation teams within online services, UK Finance has also flagged that there currently appears to be a disproportionate weighting on other parties to monitor content.<sup>416</sup> This reflects the importance of not indirectly placing a burden on external entities to detect fraud on these services, instead ensuring that online services do all they can to tackle fraud. One stakeholder has noted that the ability to request a dedicated reporting channel would also be a "useful backstop" and a "failsafe" in the event that they are "struggling to get engagement" with an online service.<sup>417</sup>

16.217 We need to consider the detail of what a DRC recommendation should look like, in order to properly evaluate its effectiveness, costs and impact on human rights. We have considered, first, who the DRC should be available to and second, what it should involve, having regard to the evidence set out above that inconsistency and lack of clarity is leading to harms.

## Options

16.218 We considered the following options.

### Who the DRC should be available to:

- a) **Option 1:** specified public bodies with expertise in identifying fraud would be eligible to use a DRC.
- b) **Option 2:** specified public and commercial entities with expertise in identifying fraud would be eligible to use a DRC.

### What should establishing a DRC involve:

- a) **Option 1:** Ofcom should make recommendations about what DRCs should be like.
- b) **Option 2:** Services should consult trusted flaggers on their arrangements for DRCs.

---

<sup>411</sup> City of London Police, 2023. Email exchange with representatives from Action Fraud on 5 September 2023.

<sup>412</sup> [CONFIDENTIAL ✕].

<sup>413</sup> [CONFIDENTIAL ✕].

<sup>414</sup> [CONFIDENTIAL ✕].

<sup>415</sup> [UK Finance](#) response to Ofcom 2022 Illegal Harms Call for Evidence, p.3.

<sup>416</sup> [UK Finance](#) response to Ofcom 2022 Illegal Harms Call for Evidence, p.10.

<sup>417</sup> [CONFIDENTIAL ✕].



*Who the DRC should be available to*

- 16.219 **Option 1:** We consider the proposed measure should at least cover the following public bodies. The **City of London Police**<sup>418</sup> is the national lead police force for fraud and cyber security, whilst the **National Economic Crime Centre**<sup>419</sup> 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.
- 16.220 The **National Cyber Security Centre** operates an online reporting portal for organisations and individuals to report scam website links and URLs<sup>420</sup>, many of which are also likely to be shared or promoted by fraudsters via user-generated content posted on online services.
- 16.221 The **Dedicated Card and Payment Crime Unit** (a joint partnership between law enforcement and the banking sector) partnered with several social media platforms to identify accounts that featured posts linked to payment crime. During the first 6 months of 2020, this partnership saw over 575 social media accounts associated with fraudulent activity taken down.<sup>421</sup>
- 16.222 The **Financial Conduct Authority** has useful insights to share with online services in relation to investment and financial promotions scams. As set out in our draft Illegal Content Judgment Guidance annex A6, paragraph 59, the FCA is also in a position to provide important information and expertise on relating to certain financial-services-related priority offences which would improve the ability for services to detect and remove opportunities for fraud to take place online.
- 16.223 In addition, 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).
- 16.224 **Option 2:** We recognise that commercial entities, such as FCA regulated financial institutions, are also at the forefront of tackling and preventing fraud. The banking and finance sector holds valuable contextual information and intelligence that can go a long way towards supporting online services to understand how criminals are targeting users online. The unique investigative capacity and expertise of the financial sector, the significant resources at their disposal, and their powerful commercial incentives to mitigate fraud, mean that the information they collectively submit is likely to be well-evidenced and useful for providers. As set out above, the financial sector has called for better engagement with online service providers and more effective reporting routes.
- 16.225 However, the sector is very large.<sup>422</sup> Even before we come to consider costs, which would be likely to be significant, we have a concern that the measure may be ineffective at reducing harm if it is made available to too many organisations at once. There would be a risk that services would end up engaging with so many complaints and of such varying quality that the DRC did not help them to prioritise effectively, together with an increased risk of

---

<sup>418</sup> Which includes Action Fraud, the National Fraud Intelligence Bureau (NFIB) and the National Economic Crime Victim Care Unit (NECVCU).

<sup>419</sup> The NECC sits within the NCA.

<sup>420</sup> NCSC. [Report a Scam Website](#), [accessed 8 September 2023].

<sup>421</sup> UK Finance,. [DCPCU prevents £12.5 million of fraud in the first half of 2020](#), [accessed 8 September 2023].

<sup>422</sup> For example, many tens of thousands of entities are regulated by the FCA.

security breaches and/or malicious reporting (for example, competitors reporting one another). In addition, commercial entities are not subject to the legal duties relating to fairness and human rights which bind public entities. For the time being, we therefore do not consider it appropriate to propose recommendations for commercial entities to have access to DRCs, though this is an area on which we intend to do further work in future.

16.226 Overall, we provisionally consider Option 1 is the better option. However, it would remain open for services to allow other organisations to use their DRC, where they consider this will improve safety for users.

### What should establishing a DRC involve?

16.227 A DRC is, in essence, simply a means for trusted flaggers to communicate with services. In practical terms, a DRC may for example take the form of a web portal, an inbox, a secure web-link, or other digital interfaces that enable a trusted flagger to securely submit information to the provider of the service. The contact information for a DRC would not be publicly available, as only organisations with appropriate expertise and intelligence would be expected to have access to it.

16.228 **Option 1:** If Ofcom were to make detailed recommendations about what a DRC should comprise, services would have an incentive to adopt them in order to be in a safe harbour. There would therefore be likely to be some industry movement towards standardisation of DRC processes, improving the ease of reporting for the organisations concerned. Greater ease may, all else being equal, lead to better outcomes for users if it meant that the reporting entity had more time for identifying and investigating possible frauds.

16.229 However, services are very different from one another. Ofcom's detailed recommendations may not meet the needs of every service or be proportionate in the light of their own systems and processes. They would also risk disincentivising innovation. At this stage, we provisionally consider that it should be for services to determine what works best for their service.

16.230 **Option 2:** However, leaving the arrangements wholly to services creates a risk of inconsistent outcomes and a continuation of the harms identified above. At the very least, trusted flaggers which would be entitled to use a DRC need some way of knowing who within the service they should contact to set one up.

16.231 In addition, if services spoke to trusted flaggers about their arrangements, trusted flaggers would have an opportunity to share their learnings on best practice and areas where difficulties have arisen in the past.

16.232 We considered whether it would be appropriate to ask services to consult trusted flaggers before establishing their DRC. However, the timing of this gave us pause. If a service waited until it had been approached to establish a DRC before attempting to consult all trusted flaggers on the arrangements, it could lead to delays in establishing the DRC and harm to users. But consulting on a DRC none of them wished to use would be unnecessary. On balance, we consider that a more proportionate means to secure that services consider the needs of trusted flaggers appropriately would be to recommend that they commit to engage with a trusted flagger to understand its needs with respect to the dedicated reporting channel, and to periodically seek feedback from all participating trusted flaggers (every 2 years) on any reasonable adjustments/improvements that might be made to the DRC's operation.

16.233 We therefore considered whether to recommend that:

- The service should publish a clear and accessible policy on its processes relating to the establishment of dedicated reporting arrangements for trusted flaggers (i.e., the entities listed above), covering any relevant procedural matters. This policy should include a commitment from the service to engage with a trusted flagger to understand its needs with respect to the dedicated reporting channel.
- If a request is made in accordance with the policy by any trusted flagger, the service should establish and maintain a dedicated reporting channel for fraud.
- At least every two years, the service should seek feedback from the trusted flaggers with which it has made such arrangements, on whether any reasonable adjustments or improvements might be made to the operation of the dedicated reporting channel.
- Complaints from trusted flaggers received through the dedicated reporting channel relating to specific content should be handed in accordance with Recommendations 4A to 4E of the draft Code. Services should ensure that complaints received through the dedicated reporting channel relating to other matters are handled as if they were relevant complaints, in accordance with Recommendation A5.H (on appropriate action for all other relevant complaints) and, where applicable, Recommendation 3E on tracking evidence of new and increasing illegal harm.

## Effectiveness

16.234 The establishment of DRCs with trusted flaggers would go a long way towards addressing the current challenges with reporting fraud, specifically the need for clearer, more effective communication channels, with consistent engagement. A recommendation in Codes that services should establish DRCs with trusted flaggers would be a basis for the entities we identify above to engage with services more effectively. Services would be incentivised by the risk of regulatory action against them to set up appropriate processes and take feedback from trusted flaggers on them.

16.235 In turn, this would be likely to improve outcomes for users. The trusted flaggers we are considering between them have significant expertise and intelligence on a wide range of fraud and we provisionally consider that establishing a DRC for their use would be of significant benefits in terms of protecting users against fraud. The trusted flaggers we have identified are well placed to accurately identify fraudulent content earlier than it would otherwise be discovered. This supports services in taking such content down more quickly, lowering the risk of users being exposed to it.

16.236 We have engaged with these organisations in the lead-up to the publication of this document and consider it likely that they would use the DRC.

16.237 We would expect complaints about illegal content from a trusted flagger to be handled appropriately as a matter of compliance with the safety duty (as these complaints are not necessarily covered by the complaints handling duty).

16.238 As set out in Chapter 12 on U2U Content Moderation and Chapter 13 on Search Moderation, the proposed approach is to leave services with flexibility as to how they design their content moderation systems (rather than being prescriptive). Instead, these chapters we set out the factors they should have regard to when considering how to design their systems.

Therefore, in accordance with these chapters, we consider that there would be risks of perverse outcomes and greater harm overall to users were we to recommend that services' content moderation functions should always prioritise for review, content reported via a DRC. However, the likelihood that such complaints will correctly identify illegal content, together with the likelihood that expert entities will tend to focus their work on the most widespread and serious harms, means that as set out in Chapter 12 U2U content moderation prioritisation paragraph 12.135, we consider that that services should have regard to whether a report came from a trusted flagger in its prioritization process.

16.239 Complaints from trusted flaggers may also be useful to identify emerging practices and harms more generally. In Chapter 8 paragraph 8.140 (governance and accountability), we are proposing that services should monitor emerging harms. Complaints made via DRCs could be a valuable input into that process.

16.240 Overall, we consider that recommending that services should establish a DRC would contribute to online services making better, more accurate and more timely content moderation decisions relating to fraudulent content, which in turn would reduce the risk of users encountering fraudulent content and becoming victims of fraud.

## Costs and risks

16.241 This measure would involve a service developing a policy on its processes relating to dedicated reporting arrangements. If a request was made from a relevant entity, the service would then need to engage with that entity and design and implement the direct reporting channel. These costs would not vary much with the size of the service and will largely be a one-off cost, though there will be some ongoing costs to maintain the direct reporting channel. At this stage, we do not have detailed evidence of the nature and scale of these costs.

16.242 We are proposing to recommend that 7 trusted flaggers should be eligible to establish DRCs. This creates an upper limit on the number of set up requests a service could receive. The entities concerned are all public bodies with an incentive to seek consistency in reporting processes which should tend to limit the likelihood of processes throwing up novel questions at least to a degree, which reduces the need to spend resources considering them. This is relevant for costs as consistent reporting processes should result in cost efficiency when developing and maintaining the reporting function. Nonetheless, we would still expect services to have regard to proposals from a trusted flagger with respect to the operation of the DRC.

16.243 As noted, there would also be ongoing costs in maintaining the DRC, managing the reporting process and dealing with reports from relevant trusted flaggers. These costs are likely to vary depending on how many relevant trusted flaggers a service links with and the volume and complexity of reports, which would themselves vary depending on the volume and seriousness of fraud on the service.

16.244 As we expect DRC reports to be well targeted at fraudulent content, these ongoing costs should be proportionate in line with benefits to users. Therefore, where services had high ongoing costs resulting from a large number of reports from relevant trusted flaggers, there would also probably be correspondingly high benefits for users from a reduction in their exposure to fraud-related content. Conversely, services that comply with this measure but have low volumes of such reports would be unlikely to have high on-going costs.

16.245 We note that some large services, for example YouTube, already have a comparable process in place.<sup>423</sup> Google (and YouTube’s) “Priority Flagger” programme provides channels for participating organisations to notify Google of potentially harmful issues on their products and services that violate their policies and Community Guidelines. Google refers to this as a “dedicated intake channel”.<sup>424</sup> Multiple services, including Snap have trusted flagger relationships with the Dedicated Card and Payment Crime Unit.<sup>425</sup> Services that are also subject to the relevant part of the EU’s Digital Services Act will already be required to establish trusted flagger schemes. Where there are synergies between that requirement and our proposals, then additional costs will be less for services subject to both. The trusted flagger scheme in the DSA does not apply to small and microbusinesses unless they are very large Online Platforms.<sup>426</sup>

## Rights impacts

### Freedom of expression impacts

16.246 We consider the impact of this measure on users’ freedom of expression to be minimal as it does not require platforms to take down content. This measure focuses on the detection process rather than enforcement. Service providers will retain discretion to decide whether a submitted report from a trusted flagger provides sufficient evidence and context to justify removal.

### Privacy impacts

16.247 The trusted flaggers submitting complaints to services would be subject to their own obligations under data protection and privacy legislation, and would be able to do so only if satisfied that they were acting lawfully. Services themselves will remain subject to those laws.

## Who the measure would apply to

16.248 As set out above, establishing a DRC would involve both set-up costs and ongoing maintenance and operational costs for the service. At this stage, we do not have detailed evidence of the nature and scale of these costs. Our provisional view is that the very significant scale and harm of online fraud is such that for some services, even high costs are proportionate.

16.249 Engaging with a DRC also creates costs for the trusted flaggers which use it. It is not likely that they would have the time or willingness to set up DRCs with every regulated service – they, too, prioritise their work. Imposing the costs on services of establishing a DRC policy which was unlikely ever to be used would not be proportionate.

---

<sup>423</sup> YouTube. [About the YouTube Priority Flagger Programme](#), [accessed 25 August 2023].

<sup>424</sup> Google Transparency Center, [Google’s Priority Flagger Programme](#), [accessed: 25 August 2023]; Google, [Supplementary Written Evidence to Parliament](#), page 7, [accessed 4 September 2023]; House of Commons Select Committee, 2019. [Impact of social media and screen-use on young people’s health Contents](#), paragraphs 184 – 185, [accessed 4 September 2023]. House of Commons Select Committee, 2019. [Impact of social media and screen-use on young people’s health](#), paragraphs 184 – 185.

<sup>425</sup> Sanjit Gill, 2021. [Written Evidence submitted by Snap Inc to Treasury Select Committee’s inquiry into online advertising and economic crime](#), [accessed 4 September 2023].

<sup>426</sup> Article 22 of the [EU’s Digital Services Act](#) sets the trusted flagger requirements. Article 19 exempts small and microbusinesses from this requirement.

- 16.250 At this stage, we are focusing our work on fraud. It follows that it is appropriate to apply the measure to services at risk of fraud, and we provisionally therefore consider it appropriate to recommend this measure for services that have identified as having a high or medium risk for fraud.
- 16.251 However, smaller services, including some small and microbusinesses and vertical search services may identify themselves as being at high or medium risk for fraud. We have limited information on the scale of the costs of establishing and maintaining the DRC and consider they could be significant for smaller services, which will tend to have fewer resources. Given that an unknown proportion of the costs of this measure would be fixed set up costs, we lack detailed evidence of these costs, and we do not have clear evidence of likely take up of the remedy by trusted flaggers in relation to all services at medium or high risk of fraud, there is a risk that the measure may not be proportionate for all services. We would be less worried about any ongoing costs of dealing with the reports for smaller services, as these are likely to scale with the benefits.
- 16.252 As set out above, fraud is a volume crime, with opportunistic perpetrators generally targeting services with large user bases. The more users exposed to the content, the greater the likelihood that the fraudster will be successful in deceiving someone. The largest services are likely to be able to resource the costs of establishing and maintaining DRC for fraud. That some large services already have a comparable process in place suggests that this measure is likely to be proportionate for such services. In general, the benefits will be considerable for these services because they reach a lot of users and, as set out above, this measure will lower the risk of users of such large services being exposed to fraud-related content online.
- 16.253 For all these reasons, we provisionally think it would be proportionate to recommend this measure for large services that have identified a high or medium risk of fraud in their most recent risk assessment. We recognise that this creates a risk of displacement of the harm, but note that even if it does, the harm to users overall is likely to be reduced to the extent that smaller services have lower reach. We may consider expanding this measure to more services in the future, as we further develop our understanding of the costs associated with running a DRC and the use of smaller services by fraudsters.

## Provisional conclusions

- 16.254 As set out above, there are very significant harms to UK users from fraud. Action Fraud reported £2.35bn in fraud related losses in 2021-22, whilst noting that 80% of reported fraud is cyber-enabled, and that social media and encrypted messaging services as an enabler is increasing throughout all aspects of fraud.<sup>427</sup> We believe that properly implemented and resourced DRCs would make a meaningful contribution to reducing this. The establishment of DRCs with trusted flaggers would go a long way towards addressing the current challenges with reporting fraud, specifically the need for clearer, more effective communication channels, with consistent engagement. The trusted flaggers we have identified are well placed to accurately identify fraudulent content earlier than it would otherwise be discovered. This supports services in taking such content down more quickly, lowering the risk of users being exposed to it.

---

<sup>427</sup> Action Fraud (City of London Police), [Annual Assessment Fraud Crime Trends 2021-22](#), [accessed 26 September 2023].

- 16.255 We expect DRC reports would be mainly targeted at fraudulent content, with only a small proportion of incorrect reports. Therefore, if the costs of dealing with reports is high, this is likely to mean the benefits to users from removing fraudulent content is also high. As such, we consider the ongoing costs of dealing with the reports that are identified through DRCs are likely to be proportionate to the benefits outlined.
- 16.256 For the reasons set out above, we are not proposing to recommend this measure for smaller services at this stage, even if they have identified a medium or high risk of fraud at this stage. We may consider expanding this measure to more services in the future, as we further develop our understanding of the costs associated with running a DRC and the use of smaller services by fraudsters.
- 16.257 Therefore, we propose to recommend that our Code on other duties contains the following provisions for large U2U and search services that have identified a high or medium risk of fraud in their most recent risk assessment.
- 16.258 We therefore provisionally consider that, as set out in Recommendation 5I in Annexes 7 and 8:
- The service should publish a clear and accessible policy on its processes relating to the establishment of dedicated reporting arrangements for trusted flaggers (i.e. the entities listed from paragraphs 16.219 to 16.223 above), covering any relevant procedural matters. This policy should include a commitment from the online service to engage with a trusted flagger to understand its needs with respect to the dedicated reporting channel.
  - If a request is made in accordance with the policy by any trusted flagger, the service should establish and maintain a dedicated reporting channel for fraud.
  - At least every two years, the service should seek feedback from the trusted flaggers with which it has made such arrangements, on whether any reasonable adjustments or improvements might be made to the operation of the dedicated reporting channel.
  - Complaints from trusted flaggers received through the dedicated reporting channel relating to specific content should be handled in accordance with section Recommendations 4A to 4E of the draft Code. Services should ensure that complaints received through the dedicated reporting channel relating to other matters are handled as if they were relevant complaints, in accordance with section A5.H for U2U services (on appropriate action for all other relevant complaints) and section A5.H for Search services (on appropriate action for all other relevant complaints) and, where applicable, section 3E of this Code on tracking evidence of new and increasing illegal harm.



# 17. Terms of service and publicly available statements

## What is this chapter about?

The Act requires that all U2U and search services must:

- **Include the following provisions in its ToS/PAS:** (a) how individuals are protected from illegal content, (b) information about any proactive technology used for compliance with the illegal content safety duties, and (c) policies and processes that govern the handling and resolution of relevant complaints.

This chapter covers the obligations services have regarding Terms of Service (ToS) and publicly available statements (PAS)<sup>428</sup>, and our proposals for Code measures in this area, both in relation to the provisions services should include in them (noted above) and how they can ensure they are clear and accessible for users.

## What are we proposing?

We are making the following proposals for all U2U and search services:

- **Ensure that the provisions included in their ToS/PAS are easy to find**, in that they are: clearly signposted for the general public, locatable within the ToS/PAS, laid out and formatted in a way that helps users read and understand them; written to a reading age comprehensible for the youngest person permitted to agree to them; and designed so people dependent on assistive technologies can access them.

## Why are we proposing this?

It is important that users be informed about how services treat illegal content. Based on our analysis of behavioural science literature, our understanding of best practice and findings from our work regulating VSPs, we consider that if services follow the recommendations set out above, these provisions will be clear, accessible and easy for users to digest. This will make users better able to make informed choices about what services to use, thereby reducing the risk of online harm.

## What input do we want from stakeholders?

- Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.
- Do you have any evidence, in particular on the use of prompts, to guide further work in this area?

## Introduction

---

- 17.1 There are several duties on regulated services aimed at ensuring that users of online services know how illegal content will be treated online. Services should ensure this is the case by

---

<sup>428</sup> A PAS is a statement made by a search service, available to members of the public in the UK, often detailing various information on how the service operates.

designing terms of service and publicly available statements that are easy to access and understand. Provisions about how users are protected must also be consistently applied, so users know how they will be protected from illegal content.

- 17.2 In this chapter, we set out our proposed recommendations as to how services can achieve their duties around terms of service and publicly available statements, as regards illegal content. We believe that these measures are integral to ensuring that users can find reliable and up-to-date information about safety practices on regulated services.

## Defining ‘terms of service’ and ‘publicly available statements’

- 17.3 The Act includes duties that apply in relation to:
- a) U2U services’ ‘terms of service’ (‘terms’), meaning “all documents (whatever they are called) comprising the contract for use of the service (or of part of it) by United Kingdom users”<sup>429</sup>; and
  - b) Search services’ publicly available statements (‘statements’): search services are required to produce, and make available to the public, statements setting out various pieces of information about how they operate.<sup>430</sup>
- 17.4 Combined services, which have both functionalities, are permitted to set out what would be required in a publicly available statement (where that differs) in terms of service instead.<sup>431</sup>

## Regulated services’ obligations regarding terms and statements

- 17.5 The safety duties contain several requirements relating to terms and statements in the context of illegal content. However, the Act’s duties relating to provisions in terms and statements may be grouped under three core areas:
- a) substance<sup>432</sup>;
  - b) consistency<sup>433</sup>; and
  - c) clarity and accessibility.<sup>434</sup>
- 17.6 The recommendations below deal directly with substance and clarity and accessibility. We also recognise the importance of consistently applied terms or statements (as the case may be). In our view, providers who properly implement recommendations applicable to them under the Code will necessarily do so in a way that ensures that terms or statements are applied consistently (by virtue of how those recommendations have been designed). While we do not set out a freestanding recommendation in this regard, this theme is developed in other areas of this consultation, in particular, in Chapters 12 (U2U content moderation), 13 (Search Moderation) and Chapter 16 (regarding User Reporting and Complaints, in relation to decisions).
- 17.7 There are other duties in the Act relevant to terms of service, including the ‘terms of service duties’<sup>435</sup>, and the requirement to summarise findings from the most recent illegal content

---

<sup>429</sup> Section 236 of the Online Safety Act 2023.

<sup>430</sup> The Act provides a definition for ‘publicly available’. Source: section 236 of the Online Safety Act 2023.

<sup>431</sup> Section 25 (2) (a) of the Online Safety Act 2023.

<sup>432</sup> Sections 10 (5), 10 (7), 21 (3), 27 (5), 27 (7), 32 (3) of the Online Safety Act 2023.

<sup>433</sup> Sections 10 (6) and 27 (6) of the Online Safety Act 2023.

<sup>434</sup> Sections 10 (8), 27 (8), 21 (3), 32 (3) of the Online Safety Act 2023.

<sup>435</sup> Sections 71 and 72 of the Online Safety Act 2023.

risk assessment<sup>436</sup>, for which we will consult on any proposals in later phases of our work as they are largely limited to a category of service.<sup>437</sup> We will consult on duties relating terms and statements to the protection of children in the consultation focused on the protection of children, which we plan to publish next year.

- 17.8 Beyond the duties and recommended measures in Codes, services can also see our report “What we’ve learnt about VSPs’ user policies”<sup>438</sup> which shares examples of good practice regarding user policies which we have observed while regulating Video Sharing Platforms.

## Measure 1: Substance

---

- 17.9 For the purposes of this chapter, our recommendations relate to the substance of the terms or statements, and their clarity and accessibility, in line with relevant duties in the Act.
- 17.10 The first of these, which requires for terms and statements to include certain substantive provisions, is specified in the following provisions in the Act:
- a) Section 10(5) requires U2U services to include provisions in the terms of service specifying how individuals are protected from illegal content, in respect of compliance with the duty in section 10(3) requiring that services’ systems and processes should minimise the time any illegal content is present (separately addressing terrorism content, CSEA content and other priority illegal content), and swiftly takedown any illegal content when they become aware of it;
  - b) Section 27(5) places a requirement on search services to include provisions in a statement specifying how individuals are protected from search content that is illegal content;
  - c) Sections 10(7) and 27(7) outline that, with regard to describing the proactive technology used to safeguard users, services must describe the kind of technology, when it is used, and how it works;
  - d) Sections 21(3) and 32(3) outline that services must specify the policies and processes that govern the handling and resolution of complaints of a relevant kind. The list of complaints of a relevant kind are set out in sections 21(4) and 32(4) respectively.
- 17.11 The duty described in paragraph (a) above, which applies to regulated U2U services, requires more specific information than the duty on regulated search services in paragraph (b). In particular, U2U services’ terms must address how individuals are to be protected from illegal content, with reference to the systems and processes operated by the service that (i) minimise the length of time for which priority illegal content is present and (ii) take down illegal content that the service becomes aware of (whether because the service has been alerted to it or otherwise).
- 17.12 Additional requirements apply with regard to the minimisation of the presence of priority illegal content on the service. U2U services must ensure that their terms address terrorism content, CSEA content and other priority illegal content separately. By comparison, search services must address how individuals are to be protected from search content that is illegal content more generally in their statements.

---

<sup>436</sup> Section 10 (9) of the Online Safety Act 2023.

<sup>437</sup> Section 72 (1) of the Online Safety Act 2023 will apply to all services.

<sup>438</sup> Ofcom, 2023. [Regulating Video Sharing Platforms \(VSPs\): Our first 2023 report: What we’ve learnt.](#)

- 17.13 Service providers are likely to take different approaches in tackling the risk of illegal content appearing on their service, and this will impact the content of the provisions they are required to include in their terms and statements. We recognise that these measures may overlap with services' existing approach to managing content which violates wider terms of service, whether this regards content that is illegal or not illegal.
- 17.14 The duties in paragraph (c) require that certain information is included regarding any proactive technology that the service relies on for the purposes of complying with the duties placed on services to protect users.<sup>439</sup> "Proactive technology" is defined in Annex 16.
- 17.15 The duties in paragraph (d) impose further requirements: providers must specify (in a way that is accessible to children) the policies and processes that govern the handling and resolution of certain complaints. These categories of complaint are particular to the UK's online safety regime and so would not necessarily be covered by services' existing terms or statements. Our recommendations relating to the handling and resolution of complaints are set out in Chapter 16.

## Provisional conclusion

- 17.16 All U2U services should include in their terms:
- a) provisions specifying how individuals are to be protected from illegal content, addressing:
    - i) separately for each of terrorism content, CSEA content and other priority illegal content, how the service will minimise the length of time for which any priority illegal content is present; and
    - ii) how, where the service is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way, it will swiftly take down such content.
  - b) provisions giving information about any proactive technology used for the purposes of compliance with any of the illegal content safety duties<sup>440</sup> (including the kind of technology, when it is used, and how it works);
  - c) provisions specifying the policies and processes that govern the handling and resolution of relevant complaints.<sup>441</sup>
- 17.17 All search services should include in their statement:
- a) provisions specifying how individuals are to be protected from illegal content;
  - b) provisions giving information about any proactive technology used for the purposes of compliance with any of the illegal content safety duties<sup>442</sup> (including the kind of technology, when it is used, and how it works);

---

<sup>439</sup> Sections 10 (2) and 10 (3) for U2U services, and 27 (2) and 27 (3) for search services, of the Online Safety Act 2023.

<sup>440</sup> For this purpose, "illegal content safety duties" means the duties in section 10 (2) and (3) of the Online Safety Act 2023.

<sup>441</sup> For this purpose, "relevant complaints" means those falling within section 21 (4) of the Online Safety Act 2023.

<sup>442</sup> For this purpose, "illegal content safety duties" means the duties in section 27 (2) and 27 (3) of the Online Safety Act 2023.

c) provisions specifying the policies and processes that govern the handling and resolution of relevant complaints.<sup>443</sup>

17.18 We consider the requirements set out in the duties above are sufficiently clear for services to implement without further elaboration by Ofcom. Given that our recommendation closely follows the specific requirements in the Act, we consider its impacts are as required by the Act.

17.19 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure.

## Measure 2: Clarity and accessibility

---

### Purpose of the measure

17.20 We have considered what measures to include in the Code regarding clarity and accessibility of terms of service and publicly available statements. The recommendation set out below is intended to secure compliance with the duties in the Act relating to the clarity and accessibility of the provisions set out in Measure 1.

### Options

17.21 In considering recommendations regarding clarity and accessibility, we have conducted research to identify the characteristics of clear and accessible provisions. We also analysed responses to our 2022 Illegal Harms Call for Evidence.

17.22 We considered two approaches to the requirements on regulated services. These were:

- a) Option A: recommending that services meet certain outcomes through providing clear and accessible provisions; and
- b) Option B: recommending that services follow specific design criteria for clear and accessible terms of service based on what the characteristics could look like in practice.

17.23 As part of this process, we considered whether to recommend the use of ‘prompts’ to encourage users to engage with terms and statements, as well as specific recommendations around languages which terms and statements should be available in.

### Effectiveness

17.24 We recognise that providing prescriptive steps or design criteria (option B) could offer a clearer safe harbour for services. However, these may not be effective or proportionate to the range of services in scope of the Online Safety regime, for the following reasons:

- a) Whilst there is clear evidence about the high-level factors which determine how clear and accessible terms of service are, the evidence as to what specific design choices are most conducive to clarity and accessibility is not always clear cut.
- b) Services may have different userbases with different needs or may be used in different contexts. This is particularly relevant given the heterogeneity of the services in scope of the Act. As an example, it would be inappropriate to recommend a specific reading age

---

<sup>443</sup>For this purpose, “relevant complaints” means those falling within section 32(4) of the Online Safety Act 2023.

requirement given that minimum age requirements can differ across services. Navigability and accessibility can also differ based on how the user accesses a service, for example on desktop, mobile or other type of internet-connected device.

- c) Services may have different ways of presenting information or navigating to terms. For example, what is considered a default 'homepage' could change over time. Assistive technologies may also change. This makes it difficult to suggest specific requirements around presentation or usability.
  - d) Further, we consider services will generally be best placed to judge what approach to presenting provisions is most likely to make them clear and accessible.
- 17.25 We therefore do not consider that option B offers enough flexibility to ensure that terms and statements are clear and accessible for every service in scope.
- 17.26 By comparison, asking services to achieve certain outcomes in their terms and statements (option A) would provide clarity about our broad expectations, whilst allowing for more flexibility in the steps that could be taken to meet the duty. We therefore focus the remainder of our analysis on option A.
- 17.27 Our evidence and analysis suggest that when determining whether provisions are clear and accessible the following characteristics are important.

## Findability

- 17.28 Ofcom research<sup>444</sup> found that one in seven people who had needed to access the terms of a social media website or platform were unable to do so on the most recent occasion this was the case.<sup>445</sup> Respondents to our 2022 Illegal Harms Call for Evidence also highlighted that terms and statements can be hard to locate and/or services could make them easier to find.<sup>446</sup> For example, Big Brother Watch highlighted that content moderation information can be in places that means it is 'obscured from users' view'.<sup>447</sup>
- 17.29 Specific suggestions included ensuring that terms are easily visible publicly or before sign-up is complete and thereafter, which was highlighted both by respondents from civil society and as ongoing practice by industry respondents.<sup>448</sup>

---

<sup>444</sup> The survey was conducted in March 2023 using an online interview administered to members of the YouGov Plc UK panel of 2.5 million+ individuals who have agreed to take part in surveys. The responding sample is weighted to the profile of the sample of UK adults aged 16+ to provide a representative reporting sample derived from the census. The sample size was 2,163 adults online aged 16+. The objective of the survey was to understand online adults' experience in using services online such as social media, search engines, video or adult websites and apps, and finding information about these services. Source: Ofcom, 2023. [Platform Terms and Accessibility](#) [accessed 6 September 2023].

<sup>445</sup> The question wording: Now thinking about the most recent time you needed to access the terms of service, community guidelines or any other type of policy document ('terms') of any social media website or platform...Which ONE of the following describes your experience in trying to find information from these services' terms? (Please select the option that best applies). Source: Ofcom 2023.

<sup>446</sup> [Big Brother Watch response to 2022 Illegal Harms Call for Evidence](#), page 1; [Global Partners Digital response to 2022 Illegal Harms Call for Evidence](#), page 3; [the National Society for the Prevention of Cruelty to Children response to 2022 Illegal Harms Call for Evidence](#), page 6; [Glitch response to 2022 Illegal Harms Call for Evidence](#), page 3; [Carnegie UK response to 2022 Illegal Harms Call for Evidence](#), page 5.

<sup>447</sup> [Big Brother Watch response to 2022 Illegal Harms Call for Evidence](#), page 1.

<sup>448</sup> [Glitch response to 2022 Illegal Harms Call for Evidence](#), page 3; [Carnegie UK response to 2022 Illegal Harms Call for Evidence](#), page 5; [OnlyFans response to 2022 Illegal Harms Call for Evidence](#), page 14; [Google response](#)

- 17.30 In line with our recommendations around User Reporting and Complaints (chapter 16), being able to find terms and statements is key to them being accessible. This means that they need to be intuitive to find and straightforward to reach through a small number of steps.
- 17.31 Further, users could benefit from having open and available access to provisions without signing in. This could include users who may want to read provisions before signing up to a service, or those who want to find out how to make a complaint but whose account has been restricted.

## Layout and formatting

- 17.32 Clear presentation of provisions can help users find and understand relevant information. This is illustrated in the research by the Behavioural Insights Team (a social purpose company exploring behavioural insights<sup>449</sup>) which found that using icons to illustrate key terms increased user comprehension scores by 34% compared with the control.<sup>450</sup> Research carried out by the Danish Competition and Consumer Authority found that icon summaries increased user comprehension scores by 38%.<sup>451</sup>
- 17.33 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.<sup>452</sup> This ratio was selected as an appropriate contrast for users with vision loss equivalent to approximately 20/40 vision (typical for users aged 80) and those with colour blindness. Ofcom research found that 15% of respondents have had difficulty reading text online generally because of weak contrast in colour between text and background.<sup>453</sup> Those with a health limitation were more likely to report having this difficulty.<sup>454</sup>
- 17.34 Additionally, responses to our 2022 Illegal Harms Call for Evidence highlighted several ways to ensure clear and accessible written layouts and formats. These included the use of space<sup>455</sup> or breaking content into clear sections<sup>456</sup>, large print<sup>457</sup>, and use of bullet points<sup>458</sup>

---

[to 2022 Illegal Harms Call for Evidence](#), page 18; [Meta Platforms response to 2022 Illegal Harms Call for Evidence](#), page 8.

<sup>449</sup> The Behavioural Insights Team, 2023. [Who we are](#) [accessed 18 September 2023].

<sup>450</sup> The Behavioural Insights Team, 2019. [Best practice guide: Improving consumer understanding of contractual terms and privacy policies: evidence-based actions for businesses](#), page 12 [accessed 6 September 2023]. Subsequent references throughout.

<sup>451</sup> Danish Competition and Consumer Authority, 2018. [Improving the Effectiveness of Terms and Conditions in Online Trade Competitive Markets and Consumer Welfare](#), 15, page 5 [accessed 6 September 2023]

<sup>452</sup> Web Accessibility Initiative, 2023. [Understanding SC 1.4.3: Contrast \(Minimum\) \(Level AA\)](#) [accessed 6 September 2023].

<sup>453</sup> The question wording: Now thinking about your time spent more widely online (i.e. beyond finding or reading terms)...Have you ever had difficulty reading information because of any of the reasons below? (Please select all that apply). Source: Ofcom, 2023.

<sup>454</sup> 15% overall say they have had a problem because of illegible text due to weak contrast in colour between the text and background: Within this, 21% amongst those whose day to day activities have been impacted a lot by a health problem or disability, and 19% amongst those who say are impacted a little by a health problem or disability, reported having this difficulty compared to 13% who are not impacted at all by a health problem or disability. Source: Ofcom, 2023.

<sup>455</sup> [Money and Mental Health Policy Institute response to 2022 Illegal Harms Call for Evidence](#), pages 2 and 3.

<sup>456</sup> [5Rights Foundation response to 2022 Illegal Harms Call for Evidence](#), page 8.

<sup>457</sup> [Refuge response to 2022 Illegal Harms Call for Evidence](#), page 6.

<sup>458</sup> [Global Partners Digital response to 2022 Illegal Harms Call for Evidence](#), pages 1 and 2; [Money and Mental Health Policy Institute response to 2022 Illegal Harms Call for Evidence](#), page 3.



or bold lettering<sup>459</sup> for key points. Several 2022 Illegal Harms Call for Evidence responses highlighted entirely different formats to aid accessibility of terms and statements, including graphics<sup>460</sup> and videos.<sup>461</sup>

## Language

- 17.35 Ofcom’s research illustrates that confusing language often prevents adult users from getting the information they need from terms and conditions published by online services.<sup>462</sup> 5Rights’ response to the 2022 Illegal Harms Call for Evidence highlighted research suggesting that many services popular among children and young people set out terms in legalistic documents, with readability scores requiring a university education.<sup>463</sup>
- 17.36 The Behavioural Insights Team found that simplifying a policy’s estimated reading age from a university graduate’s reading level to a 14-year old’s reading level led to 16.9% higher comprehension levels than the control for those who were educated to a GCSE level or below when tested on them, showing the benefits of language that is suitable for a wider range of users.<sup>464</sup>
- 17.37 Several respondents to the 2022 Illegal Harms Call for Evidence, including services<sup>465</sup> and organisations advocating for users<sup>466</sup> including children<sup>467</sup> and those with learning disabilities<sup>468</sup>, raised the need for provisions to be written in clear language understandable for the range of users on a service. The need to avoid jargon was also emphasised.<sup>469</sup>

---

<sup>459</sup> [5Rights Foundation response to 2022 Illegal Harms Call for Evidence](#), page 8.

<sup>460</sup> [ICO response to 2022 Illegal Harms Call for Evidence](#), page 2; [Global Partners Digital response to 2022 Illegal Harms Call for Evidence](#), page 3; [5Rights Foundation response to 2022 Illegal Harms Call for Evidence](#), page 8; [the National Society for the Prevention of Cruelty to Children response to 2022 Illegal Harms Call for Evidence](#), page 6.

<sup>461</sup> [UK Finance response to 2022 Illegal Harms Call for Evidence](#), page 8; [the National Society for the Prevention of Cruelty to Children response to 2022 Illegal Harms Call for Evidence](#), page 6; [Global Partners Digital response to 2022 Illegal Harms Call for Evidence](#), page 3; [Chayn response to 2022 Illegal Harms Call for Evidence](#), page 3; [Google response to 2022 Illegal Harms Call for Evidence](#), page 19.

<sup>462</sup> The question wording: You previously said that on at least one occasion, you were able to find the terms but could not get the information you needed from them. Why were you not able to get the information you needed from the terms? (Please select all that apply). 55% of those who responded were able to find the terms but could not get the information needed stated confusing language as a reason. \* Note small sample. This question was based on just 110 respondents. Source: Ofcom, 2023

<sup>463</sup> [5Rights Foundation response to 2022 Illegal Harms Call for Evidence](#), page 7.

<sup>464</sup> The study tested simplifying the Terms and Conditions of a peer-to-peer room sharing platform with sentences and words which were shorter on average. By doing this, they reduced the policy’s estimated reading age from a university graduate’s reading level to a 14-year old’s reading level. Source: The Behavioural Insights Team, 2019. page 30 [accessed 6 September 2023].

<sup>465</sup> [Trustpilot response to 2022 Illegal Harms Call for Evidence](#), page 18; [Roblox response to 2022 Illegal Harms Call for Evidence](#), page 4; [Meta Platforms response to 2022 Illegal Harms Call for Evidence](#), page 8; [Chayn response to 2022 Illegal Harms Call for Evidence](#), page 3.

<sup>466</sup> [Anti-Semitism Policy Trust response to 2022 Illegal Harms Call for Evidence](#), page 8, [Global Partners Digital response to 2022 Illegal Harms Call for Evidence](#), page 3.

<sup>467</sup> [5Rights Foundation response to 2022 Illegal Harms Call for Evidence](#), page 7; [the National Society for the Prevention of Cruelty to Children response to 2022 Illegal Harms Call for Evidence](#), pages 6 and 7.

<sup>468</sup> [Mencap response to 2022 Illegal Harms Call for Evidence](#), page 3.

<sup>469</sup> [Google response to 2022 Illegal Harms Call for Evidence](#), page 18; [Mencap response to 2022 Illegal Harms Call for Evidence](#), page 3; [OnlyFans response to 2022 Illegal Harms Call for Evidence](#), page 14; [5Rights Foundation response to 2022 Illegal Harms Call for Evidence](#), page 7; [Refuge response to 2022 Illegal Harms Call for Evidence](#), page 6; [Samaritans response to 2022 Illegal Harms Call for Evidence](#), page 5; [Trustpilot](#)

- 17.38 Many respondents highlighted the importance of providing terms in multiple languages, including citing specific languages or those languages in which a service is made available. <sup>470</sup>

## Usability

- 17.39 In the UK, around one in five people report having a disability. <sup>471</sup> 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 <sup>472</sup>, while screen readers make content on a screen accessible for those who are unable to see it. <sup>473</sup>
- 17.40 Provisions may not always be accessible to these users. Ofcom research found that approximately one in ten adults online have had difficulty reading text online because it was not keyboard navigable or difficult to navigate using a keyboard. The same proportion had difficulty reading text online because it was not compatible or was difficult to use with screen reading technology. <sup>474</sup>
- 17.41 Commonly, terms and statements can include links at the top or side of the page. For users with certain disabilities, being able to skip links avoids the obstacle of navigating them to access the provisions. <sup>475</sup>
- 17.42 Semantic elements (the tags used to indicate what type of text is on the page) in HTML, which is the standard markup language for webpages, can also help those using screen readers and keyboards to navigate through information presented. <sup>476</sup>
- 17.43 Multiple respondents to the 2022 Illegal Harms Call for Evidence addressed the point that certain users, including those who are disabled, may have different accessibility needs including reliance on assistive technologies. <sup>477</sup> Three organisations referenced compliance

---

[response to 2022 Illegal Harms Call for Evidence](#), page 18; [Airbnb response to 2022 Illegal Harms Call for Evidence](#), page 2.

<sup>470</sup> [Anti-Semitism Policy Trust response to 2022 Illegal Harms Call for Evidence](#), page 8; [Anti-Defamation League response to 2022 Illegal Harms Call for Evidence](#), page 7; [Business for Social Responsibility response to 2022 Illegal Harms Call for Evidence](#), page 5; [Carnegie UK response to 2022 Illegal Harms Call for Evidence](#), page 4; [Chayn response to 2022 Illegal Harms Call for Evidence](#), page 3; [Glitch response to 2022 Illegal Harms Call for Evidence](#), page 3; [Global Partners Digital response to 2022 Illegal Harms Call for Evidence](#), page 3; [Meta Platforms response to 2022 Illegal Harms Call for Evidence](#), page 8; [REPHRAIN response to 2022 Illegal Harms Call for Evidence](#) page 2; [The Oversight Board response to 2022 Illegal Harms Call for Evidence](#), page 6; [Refuge response to 2022 Illegal Harms Call for Evidence](#), page 6; [Wikimedia Foundation response to 2022 Illegal Harms Call for Evidence](#), page 4.

<sup>471</sup> Disability Information Scotland, 2018. [One Scotland](#) [accessed 6 September 2023]; Northern Ireland Statistics and Research Agency, 2022. [Main statistics for Northern Ireland Statistical bulletin: Health, disability and unpaid care](#), page 17. [accessed 6 September 2023]; Office for National Statistics, 2023. [Disability, England and Wales: Census 2021](#). [accessed 6 September 2023].

<sup>472</sup> Web Aim, 2022. [Keyboard Accessibility](#) [accessed 6 September 2023].

<sup>473</sup> Royal National Institute of Blind people, 2023. [Screen Reading Software](#) [accessed 6 September 2023].

<sup>474</sup> The question wording: Now thinking about your time spent more widely online (i.e. beyond finding or reading terms)...Have you ever had difficulty reading information because of any of the reasons below? (Please select all that apply) Source: Ofcom, 2023.

<sup>475</sup> University of Washington, Access Computing, 2023. [What is a skip navigation link?](#). [accessed 6 September 2023].

<sup>476</sup> MDN web docs, 2023. [HTML: A good basis for accessibility](#) [accessed 6 September 2023].

<sup>477</sup> [Glitch response to 2022 Illegal Harms Call for Evidence](#), page 3; [Global Partners Digital response to 2022 Illegal Harms Call for Evidence](#), page 3; [the National Society for the Prevention of Cruelty to Children response to 2022 Illegal Harms Call for Evidence](#), pages 6 and 7; [5Rights Foundation response to 2022 Illegal Harms Call for Evidence](#), page 8.

with the latest Web Content Accessibility Guidelines as a potential means for ensuring terms are accessible.<sup>478</sup> The Guidelines encourage reading sequences to be programmatically determinable, which is important for those using assistive technologies, and keyboard accessible amongst others.<sup>479</sup>

## Proposed factors

17.44 Under option A, we would therefore focus our recommendations in codes on these four factors and would set out that services should secure that their terms of service and publicly available statements are:

- a) easy to find, such that they are:
  - i) clearly signposted for the general public, regardless of whether they have signed up to or are using the service; and
  - ii) locatable within the terms of service/ publicly available statement;
- b) laid out and formatted in a way that helps users read and understand them;
- c) written to a reading age comprehensible for the youngest person permitted to agree to them; and
- d) designed for the purposes of ensuring usability for those dependent on assistive technologies, including:
  - i) keyboard navigation; and
  - ii) screen reading technology.

17.45 We have also considered whether it would be appropriate to add prescriptive recommendations around the languages in which provisions are published. We would expect services to cater to the needs of their userbase and consider that they have a strong commercial incentive to do so. However, given the large number of languages that are spoken in the UK, the fact that some services may target specific communities of language speakers, and the costs associated with translating terms into other languages, 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 and only had provisions in that language, there would not be an expectation for these to be translated into English.

17.46 We have also considered the evidence suggesting that ‘prompts’ (information provided in a brief and timely way) to users around terms of service or changes to terms of service can improve user understanding of them. Examples of the potential use of these were raised in responses to the 2022 Illegal Harms Call for Evidence, in the context both of informing users of updates to the terms of service<sup>480</sup>, as well as more bespoke ‘just in time’ notices (such as Roblox’s notification to users leaving to third party sites,<sup>481</sup> or Nextdoor’s ‘Kindness Reminder’)<sup>482</sup>, which flag relevant information to users at specific points in the user journey.

---

<sup>478</sup>[Roblox response to 2022 Illegal Harms Call for Evidence](#) page 4; [5Rights Foundation response to 2022 Illegal Harms Call for Evidence](#), page 8; Dropbox confidential response to 2022 Illegal Harms Call for Evidence.

<sup>479</sup>Web Accessibility Initiative, 2018. [Web Content Accessibility Guidelines \(WCAG\) 2.1 W3C Recommendation 05 June 2018](#) [accessed 6 September 2023].

<sup>480</sup>[eSafety Commissioner Australia response to 2022 Illegal Harms Call for Evidence](#), page 3; [Glitch response to 2022 Illegal Harms Call for Evidence](#), page 3.

<sup>481</sup>[Roblox response to 2022 Illegal Harms Call for Evidence](#), page 4.

<sup>482</sup>[Nextdoor response to 2022 Illegal Harms Call for Evidence](#), pages 3 and 10.

- 17.47 However, our evidence on prompts around the terms of service as a means of reducing harm to users is still limited. This includes evidence from Nextdoor, which reported that 34.6% of users who encountered their reminder withheld or edited comments that may have violated their community guidelines.<sup>483</sup> In a privacy context, the Behavioural Insights Team found that showing participants pop-up text box explanations of how a company would use their data as they typed each piece of information increased comprehension scores by 9% compared with the control.<sup>484</sup> On the other hand, there is mixed evidence that prompts can lead to fatigue for recipients<sup>485</sup> and so we would need to consider if the added friction for users would be justified.
- 17.48 We are making a recommendation around prompts in other parts of our consultation, for instance in Chapter 18 (Default Settings and Support for Child Users) and would like to investigate prompts in the context of terms of service in the future both to understand the different types of prompts that may be used, and the risks and benefits of each. We would like respondents to include relevant information on prompts in their response to our consultation question.

## Costs and risks

- 17.49 We now move on to assess the costs and risks associated with the option under consideration. Services which do not currently have provisions in their terms of service explaining how illegal content is actioned, the use of proactive technology and how users can seek redress from platforms will need to add them. We have not considered the costs of developing these sections as this is a direct requirement of the Act.<sup>486</sup>
- 17.50 The proposed measure would recommend services achieve outcomes intended to ensure that relevant provisions are clear and accessible. For services that would need to make changes because of this proposed measure, we have considered the costs as set out below.
- 17.51 The costs associated with this measure would depend on the length of the relevant sections of the provisions, given that the extent of information services need to include in provisions may vary between them. We do not expect these costs to vary greatly with the size of a service, though it is possible that the provisions which larger, more complex services need to include to comply with the Act are longer. Overall, the costs associated with the changes required to comply with this measure are likely to represent a higher share of revenue for smaller services, with smaller budgets, and services that permit younger users.
- 17.52 Our analysis suggests that the measures we are recommending will be effective in improving the clarity and accessibility of the provisions services will need to include in terms and statements. We consider that the costs for services in applying these recommendations will be relatively small and proportionate given the benefits to users in being able to find out important information.

## Findability

- 17.53 Services which do not already have provisions that are publicly available and easy-to-find would incur a one-off design and engineering cost to make the required user interface

---

<sup>483</sup> [Nextdoor response to 2022 Illegal Harms Call for Evidence](#), page 10.

<sup>484</sup> The Behavioural Insights Team, 2019, page 16.

<sup>485</sup> Backman, R., Bayliss, S., Moore, D., & Litchfield, I. (2017), [Clinical reminder alert fatigue in healthcare: a systematic literature review protocol using qualitative evidence](#), *Systematic reviews*, 6(1), page 1-6.

<sup>486</sup> Sections 10(5), 10(7), 27(5) and 27(7) of the Online Safety Act 2023.

changes to meet this requirement. We do not expect this would be costly and would note that for search services these provisions would already need to be publicly available given the requirement for their provisions to be in a Publicly Available Statement. For most services we would expect the one-off cost to be between £2000 and £5000 and potentially significantly less for simple services.<sup>487</sup> There would also be some smaller ongoing costs of maintaining this.

## Layout and formatting

17.54 Services may need to edit the formatting of the provisions to facilitate user understanding, such as adding icons, bullet points, subtitles, and white space. Services may also need to change the text format, size, and colour relative to the background so that the text is easy to read. The cost impact of this is mitigated through services retaining flexibility on how they choose to help users read and understand their terms, without the proposed measure making specific requirements. The total cost would depend on the extent of revisions required by services and the specific choices made to achieve the outcome. These changes are likely to be largely one-off costs, though services would also need to ensure they maintain suitable layout and formatting whenever they revise the provisions. We anticipate the costs of this would be similar to the findability considerations above, with similar inputs and an overall one-off cost of between £2000 and £5000, with some smaller ongoing costs of maintaining this.

## Language

17.55 Services would incur an additional cost in reviewing whether the provisions are expressed in language that is likely to be comprehensible to the youngest person that is permitted to agree to them. Services may have to revise the language used to comply with this measure. The total cost would depend on the extent to which the provisions need to be revised. For example, if there is a large difference in the reading age required to understand the provisions and the age of the youngest person who is permitted to agree to them, then services may have to make significant changes. Whilst making these changes would be a one-off cost, services would need to ensure that whenever they update these provisions, they retain the same comprehensibility in language used. As an example, to simplify 800 words of text from a reading age of 16 to a reading age of 13 we estimate it would take a relevant employee three days, costing the service between £500 and £1500.<sup>488</sup>

## Usability

17.56 Services might need to make one-off design changes to ensure the relevant provisions are keyboard navigable and compatible with screen reading tools. These changes are likely to be minimal and low cost. Services would face higher costs where measures are not currently taken, for example where 'skip links' need to be added and the levels of headings used in provisions are currently not labelled at all or done so incorrectly. We anticipate the one-off

---

<sup>487</sup> This is based on the assumption it would take up to 5 working days for a relevant employee to research the best ways to meet the requirements (and assuming their salary is similar to a Software Engineer) and up to 5 working days for a Software Engineer to implement the changes. We consider these estimates to be at the higher end of the range as for many services it will take less time to research and implement any changes. Annex 14 for a detailed description of our salary assumptions.

<sup>488</sup> Assuming a salary similar to the 'Professional Occupations' occupation within the ASHE data. See Annex 14 for a detailed description of our salary assumptions.

costs of this would be similar to (a) above, with similar inputs and an overall cost of between £2000 and £5000, with some smaller ongoing costs of maintaining this.

## Rights impacts

17.57 We did not identify any material impacts on the rights of users, including freedom of expression and privacy, in this measure.

## Provisional conclusion

17.58 The Act requires all U2U services to have clear and accessible terms of service, and all search services to have clear and accessible publicly available statements. Our analysis suggests there are four key areas which should be accounted for regarding how these provisions can be deemed clear and accessible: findability, layout and formatting, language, and useability. We consider 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 allows services more flexibility in how they meet the duty.

17.59 Further, we consider that the costs of implementing the measure are likely to be low for most services. There would be potentially greater costs for services which do not already have provisions in their terms and statements that are publicly accessible, easy to find, navigable and compatible with screen reading tools. However, given the benefits of ensuring that terms and statements are sufficiently clear and accessible with regards to user understanding, we consider this measure to be proportionate.

17.60 In any case, this proposed recommendation is to enable compliance with a specific requirement in the Act, which requires all services to have clear and accessible terms or statements. In view of the evidence about the importance of findability, layout and formatting, language and useability in ensuring that terms are clear and accessible, it appears unlikely that a service which did not have regard to these factors could meet its duties under the Act. Therefore we regard the costs of this measure as primarily driven by the requirements of the Act, particularly given the considerable flexibility we have given to services as to how they comply.

17.61 In line with the analysis above, we propose to recommend that our Illegal Content Codes of Practice on Terrorism, CSEA and other duties, contain this measure which requires all services to ensure that relevant provisions are:

- a) easy to find, such that they are:
  - i) clearly signposted for the general public, regardless of whether they have signed up to or are using the service; and
  - ii) locatable within the Terms of Service/ Publicly Available Statement;
- b) laid out and formatted in a way that helps users read and understand them;
- c) written to a reading age comprehensible for the youngest person permitted to agree to them; and
- d) designed for the purposes of ensuring usability for those dependent on assistive technologies, including:
  - iii) keyboard navigation; and
  - iv) screen reading technology.



# 18. U2U default settings and support for child users

## What is this chapter about?

This chapter sets out a package of measures relating to the default settings of child user accounts on U2U services, and the provision of supportive information at critical points of a child user's online experience. These aim to mitigate risks to children using a service to prevent them from encountering illegal harm, with a specific focus on grooming for the purposes of sexual abuse.

## What are we proposing?

The measures detailed below **apply to users aged under 18**.

We are making the following proposals for all U2U services which identify a high risk of grooming and all large U2U services which identify a medium risk of grooming. For now, these would only apply to the extent that a service has an existing means of identifying child users and would apply where the information available to services indicates that a user is a child. Where services are already using age assurance technologies, they should use these to determine whether someone is a child for the purposes of the protections set out below.

Where the only information they have is a user's self-declaration of how old they are, they should use this for the time being. However, our research shows that self-declaration is not an adequate form of age-assurance, as children often give inaccurate information about their age. Next year we will be making proposals about the deployment of age assurance technology on U2U services, as we consult on the measures services should take to protect children. This will propose/require higher standards of age verification for services which have children as users, and will be an important factor in making the measures recommended in this section effective.

### Default settings for children using a service

Services should implement default settings for child users ensuring that, if the service provides the relevant functionality:

- **Children using a service are not presented with prompts to expand their network of friends, or included in network expansion prompts presented to other users.**
- **Children using a service are not included in publicly visible lists of who users are connected to, and lists setting out who child users are connected to are not displayed to other users.**
- **Where services have functionality which allows users to formally connect with one another (e.g. become 'friends') they should ensure that people cannot send direct messages to children using the service without first establishing such a connection.**
- **For services with no user connection functionality, child users are provided with a means of actively confirming whether to receive a direct message from a user before it is visible to them, unless direct messaging is a necessary and time critical element of another functionality, in which case child users should be presented with a means of actively confirming before any interaction associated with that functionality begins.**
- **'Automated location information displays', which automatically create and display the location information for child users, are switched off.**



## Support for children using a service

Services should provide the following supportive information to children using a service in a timely and accessible manner, to help child users make informed choices about risk when they are:

- **seeking to disable one of the default settings recommended.** The information should assist child users to understand the implications of disabling the default, including the protections it affords.
- **responding to a request from another user to establish a formal connection.** The information should inform them of the types of interactions that this decision would enable, and the options available to take action against a user such as blocking, muting, reporting or equivalent actions.
- **receiving a direct message from another user for the first time.** The information should remind them that this is the first direct communication with that user and of the options available to take action against them. Where direct messaging is a necessary and time critical element of a service functionality, this information could be provided before a child user commences interaction associated with that functionality.
- **taking action against another user, including blocking and reporting.** The information should include the effect of the action (such as the interaction that would be restricted and whether the user would be notified), and the further options available to limit interaction or increase their safety.

## Why are we proposing this?

Child sexual abuse is a serious crime which can have a severe and lifelong impact on children and communities. Grooming involves a perpetrator communicating with a child with the intention of sexually abusing them either online or in person. It is coupled with children experiencing other forms of sexual abuse, including rape, CSAM offences and sexual exploitation. 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. The proposed measures above would make it more difficult for perpetrators to adopt these strategies and would therefore make grooming more difficult, thereby combating CSEA.

The measures we are proposing would have some one-off costs for services that do not already do this, which are likely to be in the order of the tens of thousands of pounds for small services and the hundreds of thousand pounds for large services. Given the extremely severe nature of the harm, we provisionally consider that it would be proportionate to expect services which are high risk for grooming to incur these costs irrespective of the size of service.

## What input do we want from stakeholders?

- Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.
- Are there functionalities outside of the ones listed in our proposals, that should explicitly inform users around changing default settings?
- Are there other points within the user journey where under 18s should be informed of the risk of illegal content?

## Introduction

---

### Harms or risks that these measures seek to address

- 18.1 The measures we are considering in this chapter are designed principally to combat grooming for child sexual abuse. We provide a brief description of this harm below. We give a fuller and more detailed overview of the evidence on grooming in the Register of Risks Volume 2: Chapter 6C (CSEA).
- 18.2 Grooming involves a perpetrator communicating with a child with the intention of sexually abusing them either online or in person.<sup>489</sup> Children may also experience other forms of sexual abuse offences online as part of the grooming process including sexual communications with a child and causing or inciting a child to engage in a sexual act.
- 18.3 While it is not possible to accurately determine the scale of online grooming, we understand it to be a widespread and growing issue which affects a significant number of children in the UK,<sup>490</sup> despite evidence of significant under-reporting.<sup>491</sup> The NSPCC reports that in 2021/22, 6,156 offences involving sexual communication with a child were recorded by police forces and that grooming crimes have risen by 80% in the last four years.<sup>492</sup> A US study found that 17% of participants experienced sexual solicitation as youths from adults they had chatted with online and 23% recalled a long intimate conversation with an adult stranger which could be indicative of online grooming.<sup>493</sup> Sexual abuse causes severe, sometimes life-long, harm to children’s emotional, mental and physical health and wellbeing.<sup>494</sup>
- 18.4 Although no quantification can fully capture the human cost of harm from grooming, we are aware of attempts to quantify the impact of CSA. While not an exact proxy, these could provide some illustration of the potential scale of the harm that can result from grooming behaviour and therefore show to some extent the materiality of benefits that would arise from a reduction in grooming.

---

<sup>489</sup> Multiple grooming offences are listed as priority offences in Schedule 6 of the Act. Please refer to the draft Illegal Content Judgements Guidance in Annex 10 for more information on the priority offences.

<sup>490</sup> NSPCC, 2020. [The impact of the coronavirus pandemic on child welfare: online abuse](#). [accessed 4 September 2023].

<sup>491</sup> Independent Inquiry Child Sexual Abuse (IICSA), 2020. [The Internet: Investigation Report](#) [accessed 20 September 2023]

Katz, C., Piller, S., Glucklich, T., & Matty, D. E., 2021. [“Stop Waking the Dead”: Internet Child Sexual Abuse and Perspectives on Its Disclosure](#). *Journal of Interpersonal Violence*, 36(9–10), NP5084–NP5104. [accessed 4 September 2023].

<sup>492</sup> NSPCC, 2022. [Online grooming crimes have risen by more than 80% in four years](#). [accessed 4 September 2023]

<sup>493</sup> Greene-Colozzi, E., Winters, G., Blasko, B. & Jeglic, E., 2020. [Experiences and Perceptions of Online Sexual Solicitation and Grooming of Minors. A Retrospective Report](#). *Journal of Sexual Abuse*, 29:7, 836-854. [accessed 20 September 2023] The study was of 1,133 undergraduate college students at two public institutions in the United States and asked about their experiences when under 18.

<sup>494</sup> See IICSA, 2022. [Victims and survivors’ experiences of child sexual abuse in institutional contexts in England and Wales](#), pp.104-112 of [accessed 20 September 2023]; IICSA, 2022. Part G: “The impact of child sexual abuse” in [The Report of the Independent Inquiry into Child Sexual Abuse](#), [accessed 20 September 2023]; Owens, J. N., Eakin, J. D., Hoffer, T., Muirhead, Y., & Shelton, J. L. E. 2016. [Investigative aspects of crossover offending from a sample of FBI online child sexual exploitation cases](#). *Aggression and Violent Behaviour*, 30, 3–14; Canadian Centre for Child Protection, 2017. [Survivors’ survey: Executive summary](#), pp.28-29 [accessed 20 September 2023].

- 18.5 The UK government estimated the economic and social cost of contact child sexual abuse in England and Wales. Accounting for inflation, this is approximately £101,700 per victim in 2022 prices, though the study acknowledged that this is likely to be an underestimate.<sup>495</sup> It is likely the actual cost is higher than this, as this was deliberately intended to be a conservative estimate, and it may understate the physical and emotional harms suffered by victims. This estimate also excludes the possibility of loss of life, despite the fact that we know that online child abuse can result in the death of children.<sup>496</sup>
- 18.6 There is also a risk that once a child has been groomed by an offender the abuse can extend to other children, including siblings and friends.<sup>497</sup> This implies that by reducing one instance of grooming it is possible this could then reduce the harm to more than one child.
- 18.7 Recognising the human cost of online CSEA is fundamental in the consideration of proportionality when mitigating against this harm. Ultimately the impact of online CSEA on children and communities is significant both in terms of severity and prevalence, as described more fully in the Register of Risks in Volume 2: Chapter 6C (CSEA).<sup>498</sup>
- 18.8 While the measures outlined in this chapter focus primarily on addressing the risk of grooming online, some of the measures may also address certain other kinds of illegal harms of which child users may be particularly at risk, or where a child's online activities might put other people in their lives at risk. These include stalking offences in relation to children and also harassment, abuse, and coercive and controlling behaviour offences. We discuss those harms in the context of each measure below, as relevant.

## Age of children for default setting and support measures

---

- 18.9 Our proposals below include default safety settings across a range of communication, networking and location functionalities, and the provision of information and support at certain points of the user journey. These measures are to protect child users from the risk of grooming and other kinds of illegal harms.
- 18.10 In formulating these proposed measures, we have considered what age threshold would be appropriate, namely whether they should be applied to users under the age of 16 or 18. The term 'child' captures a broad range of ages and social and cognitive development, and we are therefore conscious that any recommendations made must be effective at safeguarding children from online harms whilst not unduly restricting the online lives of older children, those between 16-18 years in particular.
- 18.11 As set out in our Register of Risks Volume 2: 6C (CSEA) paragraphs 6C6.11 to 6C6.13, the priority offences grouped under the category of 'grooming offences' feature the shared characteristic of involving an abuser developing a relationship with a child to facilitate child

---

<sup>495</sup> See Volume 2: Chapter 6C (CSEA) paragraph 6C6.31. The original study estimated that the cost per victim was £89,240 (in 2018/19 prices). UK Home Office, 2021. [The economic and social cost of contact child sexual abuse](#). [accessed 4 September 2023]

<sup>496</sup> 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. [accessed 20 September 2023]; Campbell, J. & 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*, May 23.; Yousif, N., 2022. [Amanda Todd: Dutchman sentenced for fatal cyber-stalking](#), *BBC News*, 15 October. [All accessed 04 September 2023]

<sup>497</sup> IICSA, 2020. [The Internet Investigation Report March 2020](#). [accessed 04 September 2023]

<sup>498</sup> Please see Volume 2: Chapter 6C (CSEA), paragraphs 6C6.21 to 6C6.31.

sexual abuse.<sup>499</sup> While many specific priority offences, for example meeting a child following sexual grooming<sup>500</sup> and sexual communication with a child, apply only if the child is under 16 years, other grooming offences relating to sexual exploitation, including those relating to the generation of CSAM, relate to all children up to the age of 18. We also have evidence that suggests that these offences are also committed against children in the 16-18 age range online:

- a) Perpetrators often deploy grooming tactics to obtain new CSAM. The IWF 2022 annual report highlights the rise in cases related to “self-generated indecent imagery”<sup>501</sup> of children. They received 1,266 reports of such content that contained children aged 16 and 17 in circumstances where many children had been forced, manipulated, and coerced by online perpetrators to produce this imagery.<sup>502</sup> Furthermore, the IWF reports that an increased number of 16 to 17-year-olds are using the ‘Report Remove’<sup>503</sup> service to flag images and videos that have been non-consensually re-shared. In a case example present in the annual report, an offender groomed multiple children online and uploaded their images online. The investigation identified that 70% of the images were of children aged 16 and 17.
- b) Perpetrators may use U2U services for the sexual exploitation of children for commercial gain such as to recruit, exploit and control their victims, including 16 and 17 year old children. The Global Report on trafficking of persons 2020 highlighted that many traffickers will use social media pages to recruit and build a relationship with individuals.<sup>504</sup> The US Federal Human Trafficking report 2020 also highlighted children being targeted in this way.<sup>505</sup> Grooming is a form of child sexual abuse that is interconnected to other forms of CSEA including but not limited to trafficking, CSAM offences and contact sexual abuse. In this context perpetrators are evidenced to utilise U2U services to groom children for the purposes of further sexual abuse and exploitation offline.

18.12 We therefore provisionally consider it appropriate to recommend that any measures proposed in this chapter apply to all child users under 18 years, to ensure that 16 and 17 year olds have protections against these broader CSEA harms. While this may have some impact on the online lives of older children and their right to freedom of expression, we consider that this is proportionate given that the measures proposed are set to default and can be disabled. Our rights assessment is set out in more detail from paragraph 18.66.

---

<sup>499</sup> Schedule 6 of the Online Safety Act 2023.

<sup>500</sup> 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)).

<sup>501</sup> Ofcom recognises that this term is often inadequate to address the breadth of imagery found and equally that the use of ‘self-generated’ can be perceived as blaming victims. Until there is a consensus for a better term, Ofcom will align with industry on the use of this term.

<sup>502</sup> Internet Watch Foundation, 2022. [The Annual Report 2022](#). [accessed 04/09/2023]

<sup>503</sup> As seen in: Internet Watch Foundation, 2022. [The Annual Report 2022](#). [accessed 04/09/2023] - An IWF and NSPCC tool launched in June 2021 where young people can remove sexual images or videos of themselves online.

<sup>504</sup> United Nations Office on Drugs and Crime, 2020. “Chapter V: traffickers use of the Internet” in [Global Report on Trafficking in Persons 2020](#). [accessed 4 September 2023]

<sup>505</sup> Human Trafficking Institute, 2020. [Federal Human Trafficking Report](#). [accessed 4 September 2023]

## Default settings for child accounts

---

### Harms or risks that the measure seeks to address

18.13 From paragraph 18.1 above, we have summarised the significant harm that can result from grooming. We summarise below, with more detail in the Register of Risks, how perpetrators are known to exploit certain functionalities offered on U2U services to contact and groom children. In addition to grooming, they may also be used in connection with other non-CSEA illegal harms in the ways identified below:

- a) **Network expansion functionalities:** These are operated by means of a network recommender system, and recommend other users to connect with, based on what the service knows about 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 a user has a mutual connection. As set out in our Register of Risk Volume 2: Chapter 6C (CSEA) paragraphs 6C7.72 to 6C7.73, these functionalities can play a role in facilitating grooming. For a perpetrator to groom a child, they need access to a child, or multiple children; perpetrators use network expansion functionalities to identify children to begin grooming with either a single target, or to contact hundreds of children in a “scatter gun” approach.<sup>506</sup> If perpetrators are connecting with children then they are likely to be included in expansion prompts for other children within that child’s network thus increasing the risk for that group of child users. In addition, the National Crime Agency (NCA) comments in its National Strategic Assessment on perpetrators’ use of ‘varied personalities’ to gain access to vulnerable children to abuse. This includes perpetrators posing as children to gain trust and access to them.<sup>507</sup>
- b) **Connection lists:** 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 of Risk Volume 2: Chapter 6C (CSEA) paragraphs 6C7.47 to 6C7.49, such functionalities may be exploited by those seeking to groom children for the purposes of sexual abuse. We understand that perpetrators may utilise mutual connections to increase children’s confidence in communicating with them. We also understand that blackmail is commonly used to generate CSAM imagery, which 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.<sup>508</sup>
- c) **Direct messaging functionalities:** These allow text-based exchanges between two users in an interface that cannot be viewed by other users.<sup>509</sup> As outlined in the Register of

---

<sup>506</sup> NCA, 2021. [National Strategic Assessment of Serious and Organised Crime 2021](#), page 20. [accessed 4 September 2023]

<sup>507</sup> NCA, 2021. [National Strategic Assessment of Serious and Organised Crime 2021](#), page 20. [accessed 4 September 2023]

<sup>508</sup> Joleby M, Lunde C, Landström S, Jonsson LS. [Offender strategies for engaging children in online sexual activity](#). *Child Abuse Negl.* 2021 Oct;120:105214. Epub 2021 Jul 22. PMID: 34303993. [accessed 04/09/2023]  
Kopecký, K., 2017. [Online blackmail of Czech children focused on so-called "sextortion" \(analysis of culprit and victim behaviors\)](#). *Telematics and Information*, 34 (1), pp. 11–19. [accessed 4 September 2023].

<sup>509</sup> 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

Risk Volume 2: Chapter 6C (CSEA) paragraphs 6C7.59 to 6C7.64, these functionalities may also be exploited for grooming offences, as perpetrators can develop relationships with children away from public view and parental supervision. By sending a direct message to a child, perpetrators can initiate contact and begin the grooming process in a private online space, which affords a lower likelihood of detection or platform moderation resources. In nearly three quarters of cases (74%) when children are contacted online by someone they don't know in person, this contact involves private messaging.<sup>510</sup> Direct messaging functionalities may also be exploited for other non-grooming illegal harms directed at child users. The ability to receive a direct message could put children at risk of being targeted with harassment, threats or abuse, including hate, particularly by users that a child does not know offline Volume 2: Chapter 6E (Harassment, stalking threats and abuse offences) paragraphs 6E6.67 to 6E6.68 and Volume 2: Chapter 6F (Hate offences) paragraphs 6F6.48.

- d) **Location information** may be displayed or shared on U2U services either automatically by the provider of the service through particular functionalities, for example through “live” location functionalities or the automated display of location in user profiles or shared content, or through manual input by users on shared content. As outlined at Volume 2: Chapter 6C (CSEA) paragraph 6C7.67 the display of location information could provide a perpetrator with the necessary information to build up their knowledge base of locations frequently visited by a child, such as their home, school or other local places, and ultimately enable them to physically approach the child offline which may lead to contact sexual abuse. We are particularly concerned when children's location is displayed automatically and they may not be aware of it being disclosed. We consider this to be a more substantial risk than the instances of manual sharing, which require a proactive decision from the user to share their location. Automatic location sharing functionalities could also enable children to be live tracked without their knowledge, which we consider increases the risk of a perpetrator successfully locating a child offline. The ICO has observed in the Children's Code that geolocation data, because of the ability to track the physical location of a child, is of “particular concern” for children as it risks compromising their physical safety and renders them vulnerable to sexual abuse, abduction, physical and mental abuse.<sup>511</sup> In particular, the ICO's Children's Code highlights the importance of children being aware that their location is being shared, through their requirements that geolocation sharing options be switched off by default and that child users are made aware if their location is being tracked.

- 18.14 As identified in Volume 2: Chapter 6E (Harassment, stalking, threats and abuse offences) paragraph 6E6.73 to 6E6.74 location information may also be used to commit or facilitate other kinds of illegal harms directed at children, such as stalking, threats or other abuse. Children in care are sometimes moved across the country for safeguarding reasons.<sup>512</sup> Children with location information on by default could again broadcast their location to the individuals they were moved to prevent contact with and be at risk of significant offline harm.

---

<sup>510</sup> Office for National Statistics, 2021. [Children's online behaviour in England and Wales: year ending 2020](#). [accessed 04/09/2023]

<sup>511</sup> ICO, 2022. [Age Appropriate Design: a code of practice for online services](#). [accessed 20 September 2023]. We refer to this as the 'Children's Code'.

<sup>512</sup> Department for Education, 2021. [The Children Act 1989 guidance and regulations. Volume 2: care planning, placement and case review](#). [accessed 4 September 2023]



- 18.15 In some cases, the visibility of a child’s location online will also disclose the location of a parent/carer in circumstances where that person is a victim or survivors of abuse (as outlined above), or of coercive and controlling behaviour, as set out in the Register of Risk Volume 2 Chapter 6G (Controlling or coercive behaviour) paragraphs 6G3.66 to 6G3.68.
- 18.16 Refuge reports that domestic violence perpetrators frequently seek to use online platforms to determine a survivor’s location, for example via location settings and geo-tagging functions. 19% of survivors surveyed said their location had been compromised through tech abuse, which suggests that children may also be vulnerable in this way and has implications for their physical safety. 12% of women reported their children had been subjected to online abuse by their partner or former partner. 41% of survey respondents listed location tracking as part of coercive controlling behaviour.<sup>513</sup> Where one parent has exited an abusive relationship, children with location information defaulted to “on” for certain functionalities could give away their location to a dangerous or abusive parent without meaning to, putting both them and their parent in danger.<sup>514</sup>

## Options

- 18.17 In considering how to address the risks and harms set out above, we have looked at the following options:
- a) restrictions on network expansion and connection list functionalities:
    - i) Not including child users in network expansion prompts for other users.
    - ii) Not presenting child users with network expansion prompts.
    - iii) Not including child users in the connection lists of other users.
    - iv) Not making child users’ connection lists visible to other users.
  - b) restrictions on one-to-one direct messaging functionalities:
    - v) For services with a user connection functionality: Removing the ability for non-connected users to send direct messages to child users.
    - vi) For services without a user connection functionality: Where a new direct message is exchanged with a user for the first time, presenting child users with the ability to choose if they wish to view the message or not.
    - vii) Restrictions around the automatic display of location information relating to child users. Specifically, not automatically displaying location information in shared content or profiles of children or in live location functionalities.
- 18.18 We include more detail on some of these proposals when we discuss their efficacy below.
- 18.19 We considered the option that services should change defaults to be in line with the restrictions above, but they should allow child users to switch the functionalities set out above back on again should they wish to.
- 18.20 We considered the option that services should permanently disable the features for child users rather than just changing defaults. However, we recognise that this would have substantial implications for children’s rights to freedom of expression and freedom of association. Such implications might be, for example, that child users could be

---

<sup>513</sup> Refuge, 2021. [Unsocial Spaces](#). [accessed 4 September 2023]

<sup>514</sup> Nikupeteri, A., Katz, E., and Laitinen, M., 2021. [Coercive control and technology-facilitated parental stalking in children’s and young people’s lives](#). *Journal of Gender-Based Violence* 5 (3), pp. 395-412 [Accessed 14 July 2023]



disproportionately restricted in their ability to make beneficial connections online. In addition to human rights considerations, permanently disabling any of these functionalities would restrict the ability of child users to develop an enterprise which relies on monetising content. We do not have sufficient evidence to justify this level of interference with children’s rights and online experiences at this point.

## Effectiveness

### Building on existing practice adopted by some services

18.21 To some extent, the measures we consider build and expand upon existing practice adopted by some services already, which we provisionally consider provide effective tools for mitigating and managing the risk to children online. For example, we understand that some services restrict or provide users with the option of disabling the features covered in our measures, depending on the user’s age. For example:

- a) On TikTok, the ‘suggest your account to others’ feature is, by default, turned off for users aged under 16 and needs to be actively enabled in privacy settings.<sup>515</sup>
- b) 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’.<sup>516</sup>
- c) 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.<sup>517</sup>

18.22 We discuss the effectiveness of the measures below. We first consider the benefits of changing defaults in general, then we consider each group of functionalities in turn. Finally, we consider the residual risk and overall impact on grooming.

### Effectiveness of changing default settings

18.23 We recognise that the efficacy, and therefore the benefits, of our proposed default safety measures would be reduced if children change the default settings. They may do so voluntarily, or in some cases may be pressured to do so by peers or perpetrators. This limitation is inherent in the framing of these measures as defaults, rather than permanently disabled features of a service. As outlined above, these measures seek to strike an appropriate balance between reducing the risk of grooming and other kinds of illegal harms experienced by as many children as possible, while also ensuring their ability to exercise choice and their right to freedom of expression and association online.

18.24 Even though some children could change the default, we consider that a significant number will not, and hence the considered measures would be effective at reducing the risk of grooming harms for those children. Without the considered measures, a greater number of children would therefore remain exposed to the potential harm.

---

<sup>515</sup> TikTok, 2023, [New features for teens and families on TikTok](#) [accessed 28 September 2023]

<sup>516</sup> Instagram, 2021, [Continuing to Make Instagram Safer for the Youngest Members of Our Community](#) [accessed 28 September 2023]

<sup>517</sup> SNAP, 2022, [Parent’s Guide: Snapchat’s Family Center](#) [accessed 28 September 2023]

- 18.25 When presented with pre-set courses of action or ‘defaults’ people often tend to stick with the default option.<sup>518</sup> Default settings have been shown to strongly affect behaviour across a range of different settings.<sup>519</sup> ‘Choice architecture’ is a concept describing the context in which users make decisions and how choices are presented to them. In terms of the effectiveness of different types of interventions to influence decision-making behaviour, the CMA’s analysis of evidence around defaults and other choice architecture interventions has found that changing or setting defaults is more effective at influencing consumer choices and behaviour than other types of interventions such as changing information.<sup>520</sup>
- 18.26 Several studies suggest that choice architecture could be applied in a range of online contexts to encourage user safety.<sup>521</sup> In one specific study we are also aware that in the context of privacy settings, many children may know how to change their settings, but many choose not to.<sup>522</sup> We recognise that the literature about the application and effectiveness of choice architecture to online safety is evolving and developing, particularly as regards the impacts on safety outcomes in real-world settings and consider that it is an important area to keep under review and engage with platforms about.
- 18.27 Given we expect defaults to influence the behaviour of some children, it follows that if services are designed to default to safer settings for child users, this would tend to make children overall safer in their online experiences. We also consider that residual risk associated with some children switching off the default settings would be mitigated through our proposal to recommend that child users are provided support at the point of doing so, as described in the second measure in this chapter.

## Network expansion and connection list functionalities

- 18.28 As set out above, we understand that perpetrators deploy various techniques to approach children, including (but not limited to) sending “scatter gun” connection requests to large volumes of children and infiltrating children’s online friendship groups. We provisionally consider that defaulting network expansion and connection functionalities to ‘off’ for child users would make it materially more difficult for perpetrators to deploy these strategies:
- a) Ensuring that child users are not included in network expansion prompts and connection lists would make it harder for offenders to passively identify and connect with child users they do not know, in either a targeted or scatter gun approach. Since we understand that perpetrators may use mutual connections to generate trust, not disclosing mutual connections would reduce the risk of children connecting with

---

<sup>518</sup> Thaler, R. H., Sunstein, C. R., & Balz, J. P., 2013, Choice architecture. In E. Shafir (Ed.), *The behavioral foundations of public policy* (pp. 428-439). Princeton, NJ: Princeton University Press.

<sup>519</sup> Jachimowicz, J., Duncan, S., Weber, E., & Johnson, E., 2019. [When and why defaults influence decisions: A meta-analysis of default effects](#). *Behavioural Public Policy*, 3(2), pp. 159-186 [accessed 29 August 2023]

<sup>520</sup> Competition and Markets Authority, 2022. [Evidence Review of Online Choice Architecture and Consumer and Competition Harm](#). [accessed 29 August 2023].

<sup>521</sup> Thaler, R. H., Sunstein, C. R., & Balz, J. P., 2013, Choice architecture. In E. Shafir (Ed.), *The behavioral foundations of public policy* (pp. 428-439). Princeton, NJ: Princeton University Press; Acquisti et al. 2017. [Nudges for Privacy and Security: Understanding and Assisting Users’ Choices Online](#). *ACM Comput. Surv.* 50, 3, Article 44 (August 2017)., Available at SSRN: <https://ssrn.com/abstract=2859227> or <http://dx.doi.org/10.2139/ssrn.2859227> ; Gold, N., Lin, Y., Ashcroft, R., & Osman, M., 2023. [‘Better off, as judged by themselves’: Do people support nudges as a method to change their own behavior?](#) *Behavioural Public Policy*, 7(1), pp. 25-54 [accessed 20 September 2023]

<sup>522</sup> Livingstone, S., Stoilova, M., Nandagiri, R., 2019. [Children’s data and privacy online: Growing up in a digital age](#), pp.24 [accessed 05 September 2023]

perpetrators who may have connections in common. This should reduce the speed and volume at which perpetrators can contact children and reduce the ease with which they can identify children on the platform more generally, thereby reducing the amount of grooming which is initiated online and in turn reducing the amount of resulting sexual abuse which occurs relative to a counter-factual in which the measures were not in place.

- b) Not prompting child users to expand their own networks would likely decrease the risk of children inadvertently connecting with potential perpetrators. This should reduce the risk of children connecting with perpetrators who, as evidenced in paragraph 18.13(a,b), may have created a false persona or established many mutual connections with a child. This may give a child a false sense of security and mislead them into sending a connection request to a potential perpetrator. A reduction in the number of connections a child has with these types of users reduces the likelihood of grooming being initiated on the platform and consequently would be expected to lead to a lower level of harm.
- c) Similarly, ensuring child users' connection lists are not visible to others would make it harder for perpetrators to infiltrate children's online networks. This may stop perpetrators from using a child's connected users to build trust to begin or to further the grooming process, or from blackmailing children they have abused online into sending them further abuse images. More broadly and similarly to network expansion functionalities, this should reduce the ease with which perpetrators may identify child targets on a service.

## Direct messaging functionalities

- 18.29 As explained above, perpetrators often exploit direct messaging functionalities to initiate contact with child users and begin the grooming process in a private online space, reducing the likelihood of being seen by other users or platform moderation resources.
- 18.30 By introducing friction in the process of being able to send child users direct messages, we seek to reduce the risk of children from receiving unsolicited messages from perpetrators (including sexual images).<sup>523</sup> It is expected to make it harder for potential offenders to establish communication with a child, whether that be with a view to committing a grooming offence, or to otherwise engage in communication that would expose the child to a risk of other relevant kinds of illegal harms such as threats, harassment and abuse (including hate). This should reduce not only the amount of online grooming, but also online and offline child sexual abuse, including the production of CSAM which is commonly involved in the context of a grooming relationship, compared to a counter-factual in which this measure was not in place.
- 18.31 We provisionally consider that it might be appropriate to recommend two different approaches to implementing the default setting for direct messaging, to account for services that have a formal connection functionality and those that do not.
- 18.32 For services with a formal connection functionality, we considered the option that the default should be set to remove the ability for non-connected users to send direct messages to child users. This, we expect, would reduce the ability of potential offenders to establish communication with a child. Within this chapter we consider that a non-connected user is defined as a user whose connection has not been validated by the child user that would

---

<sup>523</sup> Unsolicited messages refer to message that are unwanted; they may or may not include sexualised content.

receive the message. A validation can either be through that child initiating the connection<sup>524</sup> or by the child confirming a connection initiated by another user.<sup>525</sup> We consider that this approach to user connections strikes an appropriate balance between ensuring that children are protected from messages from unknown users, but also ensures that children are able to communicate with other users more easily when they choose to.

- 18.33 For services without a formal connection functionality, we are conscious that it would not be proportionate to prevent non-connected users from initiating direct communication with child users as, in practical terms, that would likely involve disabling these functionalities for child users entirely, by default. This would have significant consequences not only for the freedom of expression rights of children, but also for the provider of the service.
- 18.34 We nonetheless consider it important to address the risk of harm to child users on services without a formal connection functionality and have considered the merits of recommending that those services provide child users with a means of actively choosing to view or not view the direct message from another user before the content becomes visible. We have also considered whether it would be appropriate to allow an alternative approach to meeting this requirement in circumstances where the impact of requiring a confirmation would be noticeably detrimental to the purpose of the messaging exchange, such as where receiving a message is a necessary and time critical element of another service functionality that a child user is engaging with. A potential example might be within a gaming environment when swift one-to-one messaging with other unconnected users is integral to the gameplay. Under these circumstances, services may wish to provide an alternative option for users to confirm they would like to receive one-to-one messages. For example, this could be before starting a game.
- 18.35 On those services, it is often easier for users to begin to privately message other users for the first time. We have evidence of grooming perpetrators taking advantage of services with open channels of communication that do not require a formal user connection (such as some online gaming services) to establish contact with child users that they do not know, as set out in our Register of Risk Volume 2: Chapter 6C (CSEA) paragraphs 6C7.23, 6C7.27 to 6C7.28. These conversations can then often occur in spaces that are less moderated.
- 18.36 The introduction of a prompt before the message is visible to the user would create friction at a critical point of a child engaging with a potential perpetrator, without preventing these communications entirely. It also provides child users with a greater level of choice about what content they receive in these online spaces.
- 18.37 We don't consider it appropriate to specify the precise wording and technical presentation of the information. However, to ensure that this measure is effective, the prompt must be displayed before the message is visible to child user and it must provide the child user with a means to choose whether or not they view the message.
- 18.38 In terms of presentation, we recognise that there may be a range of appropriate approaches to create this friction. Examples could include messages from non-connected users being filtered to a bespoke inbox, enabling the child user to make a choice whether to view them or not; or an immediate 'pop up' informing the child user that another user has, or wishes

---

<sup>524</sup> For example, by sending a 'friend request' or following another user.

<sup>525</sup> For example by confirming a friend request sent by another user or reciprocating a 'follow'.

to, send them a one-to-one message, including a mechanism for the child user to make a choice to view it.

## Automatic location sharing

- 18.39 Reducing the automatic display of location information about child users would reduce the risk of services being used to facilitate in-person encounters that could result in a variety of contact harms to children, such as CSEA offences, as well as stalking, harassment and abuse.
- 18.40 We are concerned that children are less likely to have the awareness of the risks associated with displaying their location information. Ensuring that the functionalities which automatically display location information in shared content, profiles, or other live location functionalities are not active by default would mean that it was less likely that children accidentally expose themselves and others to risks related to the display of their location online.
- 18.41 Recognising the availability of this information online as a potential route of exploitation for children, various regulatory instruments in the UK recommend restricting geolocation functionalities for child users. For example, the Government’s interim code of practice on online CSEA makes multiple references to defaults for children, including their geolocation being set to off by default and having default settings that are appropriate to the actual age of their users.<sup>526</sup> As outlined above, the ICO Children’s Code also seeks to address this risk.<sup>527</sup>
- 18.42 We are aware of existing practice in industry, including that X (formerly known as Twitter) has location off by default for all users on posts, and Snapchat has location-sharing off by default for all users. Users have the option to decide to share it on the “Snap Map” with friends—but never with strangers.<sup>528</sup>
- 18.43 We note that child users may use functionalities which enable them to manually (or otherwise voluntarily) display their location information via inputs on shared content or elsewhere on their user profile, or through the content itself. While we recognise that this presents a risk to children across relevant illegal harms identified above, such as grooming, CSEA, stalking and abuse, we consider that it would be most effective to focus any potential measures on automatically displayed location information. As noted above in paragraph 18.13 (d), we think there is a more considerable risk with automated location sharing, as child users may not be aware their location is being shared, as opposed to manual sharing of location where a voluntary choice is made. Additionally, requiring services to monitor all content from child users for manually displayed location information would unduly restrict the child’s rights to freedom of expression and privacy and is likely to be difficult for some services to implement, and in some cases may involve the use of costly proactive technology. On the other hand, we believe that adjusting default settings around automatic location information functionalities is likely to be more easily implemented by the majority of services and would address the risk identified about children sharing their location without being aware.

---

<sup>526</sup> Home Office, 2020. [Interim code of practice on online child sexual exploitation and abuse](#). [accessed 20 April 2023]

<sup>527</sup> ICO, 2022. Standard 10 in [Age Appropriate Design: a code of practice for online services](#). [accessed 20 September 2023].

<sup>528</sup> Snapchat. [Snapchat Support](#). [accessed 11/10/2023]

## Overall impact on grooming and residual risk

- 18.44 It is important that in the introduction of frictions into the grooming process to safeguard children online that they are still able to experience the positive benefits of being online. We have therefore designed them in such a way to ensure that children still have pathways to identify and make beneficial connections with other users, including children and adults who they know offline and to interact with users previously unknown to them.
- 18.45 It would still be possible with the introduction of the measures we have considered for users to find and connect with children on the service (and vice versa), including through ‘on-platform’ search functionalities. Children would also still be able to publicly communicate with users previously unknown to them, this can include engaging with other users’ uploaded content, engaging in forums, interest pages or games and interactions by other users on their content. This communication can move into a private space when the child user accepts a formal connection or agrees to the contact.
- 18.46 We recognise that, by not seeking to restrict any of these points of connection and communication, child users may remain at risk from certain illegal harms, including grooming. However, in striking an appropriate balance, we did not consider it proportionate to disable the ability for children to privately communicate and interact with others online.
- 18.47 We nonetheless consider our option of default settings would materially decrease the risk of child users being targeted by grooming perpetrators, as they would add friction and disrupt the current ease with which this may be done through network expansion prompts, connection lists and direct messaging.

## Costs and risks

- 18.48 The options we have considered have a number of costs, which we discuss below in turn:
- a) Direct costs of modifying services;
  - b) Indirect costs to children as a result of lost functionalities;
  - c) Indirect costs to adults as a result of friction associated with contacting children;
  - d) Indirect costs to services resulting from lost revenue.

### *Direct costs of modifying services*

- 18.49 We understand that the direct costs of modifying a service in line with the proposals would depend on two main factors: 1) the engineering costs to enable a service to modify their functions to ensure they can be switched off and defaults are set appropriately for child users; and 2) the review and overhead costs associated with any change which impacts a service’s user.
- 18.50 We consider that the engineering changes required to implement the measures would vary significantly for different services, depending on their existing system. This means there is a wide range of potential engineering costs associated with the measures. The costs could effectively be minimal if a service already provides options which allow users to choose to limit their appearance in network expansion suggestions and the visibility of their connections – in which case it would be a question of ensuring the function is defaulted on for child users.<sup>529</sup> However, the costs could be more material for services that do not

---

<sup>529</sup> As we have set out above, a number of services offer such options.



currently offer users the option of turning off the functionalities we are targeting or if they do not have a user connection functionality and need to develop a system that allows users to confirm whether they would like to receive a direct message. In such cases, we think that it is likely that services would not only need to upgrade the backend of their websites (e.g. databases and data storage), but would also need to upgrade their user interface.

- 18.51 Depending on the current structure and user interface service, any engineering change could involve graphic designers, web designers, content teams and developers and engineers. The costs associated with these modifications would likely be lower if a service is using an off the shelf tool like WordPress to build and maintain its site and higher if the service needs to modify the underlying code and site infrastructure. The cost of upgrading those aspects will also depend on the current structure of their systems. For example, the incorporation of a new functionality is likely to be significantly cheaper if the existing systems already incorporate some level of privacy design (eg, privacy options within existing backend databases) or is designed on a modular basis which separates individual functions into independent programming modules.
- 18.52 We consider an appropriate range that captures the order of magnitude of the upfront engineering costs is to assume that to implement all considered measures related to default settings for child users, services are likely to require approximately 2 to 12 months of staff resources, made up of both software engineering time and other professionals (eg, project management). We assume that this results in one-off cost of approximately £10,000 to £100,000.<sup>530</sup> The variation in costs may be partly driven by size differences but also the extent to which services need to add functionalities and/or change the user interface, or whether they just need to change the default settings of existing functionalities.
- 18.53 In addition to the engineering costs, we would expect there to be additional overhead and coordination costs associated with any change to the backend or frontend systems of a service. We expect these costs to be substantial in a large platform, where we understand significant review, communication and legal processes would need to be followed in order to implement some of the measures. We consider this includes communication of any change through a large company and the time associated with decision making processes.
- 18.54 We consider the overhead and coordination costs would be largely correlated with the size of a company. A large company with 1000s of staff is likely to have significant costs associated with these processes, whereas these costs would be much smaller for companies with a small number of staff. We consider an appropriate range that captures the order of magnitude of the operational and governance costs of these measures is to assume a service will require between 0 and 24 months of labour from a range of professionals. Using our standard assumptions results in a one-off cost of £0 to £200,000,<sup>531</sup> with the costs lying towards the lower end of the range for smaller services and towards the higher end of the range for larger services.
- 18.55 In addition to the one-off costs associated with these measures, we expect there to be a small amount of ongoing costs to review and monitor the measures and ensure the technical functionality of the measure is operating as intended. We consider an appropriate estimate

---

<sup>530</sup> Based on our standard assumptions for labour costs set out in Annex 14.

<sup>531</sup> Based on our standard assumptions for labour costs set out in Annex 14.



for ongoing costs would be approximately 25% of the original one-off cost on an annual basis.<sup>532</sup>

18.56 Table [18.1] below summarises our assumptions for the direct costs. These estimates are for the implementation of all grooming measures related to default settings as described in this section. We have not estimated the cost of individual measures but we consider the magnitude of costs to be similar across all the measures. Our estimate of costs includes the impact of efficiency savings that arise when multiple measures are implemented at the same time.

**Table 18.1 – Summary of direct cost estimates**

	Low Estimate	High Estimate
One-off cost - Engineering	10,000	100,000
One-off cost - Overhead and coordination	0	200,000
<b>Total one-off costs</b>	<b>10,000</b>	<b>300,000</b>
Total ongoing costs	2,500	75,000

Source: Ofcom analysis

#### Indirect costs to children

18.57 These measures may have an adverse impact on children’s ability to share and receive information and make new friends online. This could be a material impact as many children meet new friends online.<sup>533</sup> Less frequently, it could also affect the ability of children to monetise their platforms by attracting users to their content.

18.58 Safeguarding children online must be carefully balanced against ensuring they have access to positive connections and experiences online. While our measures impact network expansion functionalities, children would still be able to make new friends through, for example, making and receiving comments on either their own content or content uploaded by platforms and other users, and also through communication with other users in more public spaces such as in forums, on ‘pages’ or group chats. We have designed the measures we are proposing in such a way as to introduce friction into the grooming process at certain points where evidence set out in the Register of Risks Volume 2: Chapter 6 (CSEA) paragraphs 6C7.20 to 6C7.73 points to specific functionalities as presenting ‘key risks’ of children experiencing harm from grooming, while not removing other ways of interacting. We also considered the importance giving children control over these settings to make ongoing choices about their own online experiences. The ability of children to change the defaults if they wish means they can mitigate these indirect costs to some extent.

#### Indirect costs to adults

18.59 The measures could make it harder for adults and children to connect with children that they know. Instead of clicking on friends lists or relying on network expansion prompts to

<sup>532</sup> Based on our standard assumptions for ongoing maintenance of software changes set out in Annex 14.

<sup>533</sup> For example, a majority of US teenagers have made new friends online. Lenhart, A., 2015. [Teens, Technology and Friendships](#). Pew Research Center [accessed 5 September 2023]

find children they want to link with or contact, they would have to search directly for these children and send them a connection request. Once they have connected with them, they could interact with them as normal. We consider that in these instances the adults connecting to these child users are likely to consider this alternative connection pathway proportionate.

- 18.60 By design, the measures would make it harder for adults to connect with children they do not know online. To the extent that they wish to make such connections for legitimate reasons, this might impose some indirect costs.

### Indirect costs resulting from loss of revenue

- 18.61 It is possible that the measures in question could result in lower activity on sites, thereby resulting in services losing some revenue.
- 18.62 Conceptually, it is important to distinguish between (i) reductions in revenue resulting from less illegal activity taking place; and (ii) reductions in revenue resulting from the measures reducing legitimate activity on websites. We are not treating (i) as a relevant cost to factor into our impact assessment. However, we recognise that the considered measures may also reduce the volume of some legitimate interactions, it is important to have regard to this when assessing the impact of the proposed measures. It is not possible to quantify the indirect costs to services which the measures would have. However, the fact that, as we explained above, a number of services currently implement measures similar to those we are proposing (as set out above, Snap and TikTok implement similar measures by default, whereas Instagram gives children the option of selecting similar protections), suggests that the indirect costs to services are likely to be manageable.
- 18.63 There is also a small risk that indirect costs from these measures may be particularly felt by newer smaller services that are looking to grow. Growth of online services often depends on the presence of network effects, this may limit to some extent the rapid growth of user networks. Although we think this is a plausible concern, we consider it is likely to be a relatively small effect and have considered it when determining whether these measures are proportionate for smaller services.
- 18.64 Set against the indirect costs discussed above, it is also possible that the measures could have a countervailing positive effect on engagement with services. To the extent that reductions in attempts at grooming and other forms of unsolicited contact from strangers (including harmful contact such as receiving unsolicited sexual images) result in child users feeling more comfortable on a service, it is possible that they may use that service more.

## Rights impacts

### Freedom of expression and association

- 18.65 We acknowledge that these default settings could have an impact on children's and adults' rights to freedom of expression and freedom of association:
- a) The network expansion and connection list defaults would reduce the ease with which children may make connections and communicate with other users, and with which other users might encounter and explore the content produced by the child user. 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.

b) The direct messaging and location information defaults could restrict legitimate communication and engagement between children and other users on the service. This impact may be particularly acute for children using U2U services on which direct messages between non-connected users is an integral part of the operation of the service, such as during certain game play. However, we consider that the alternative approach to delivering the default setting for direct messages outlined above provides a relevant safeguard to this potential interference.

18.66 An interference with these rights must be prescribed by 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. We believe the impact is mitigated by our recommendation that the measure be implemented as a default setting that children may disable if they wish. Any residual risk to these rights is proportionate to the overall reduction in the risk of harm to children posed by perpetrators who may utilise these service features to make contact with potential grooming victims and to engage in behaviour that puts children at risk of CSEA offences. In that regard, the measures contribute to the prevention of crime and/or the protection of health or morals.

18.67 We therefore consider that the recommendation of this measure is a proportionate interference with the rights to freedom of expression and association.

## Privacy

18.68 We don't consider there to be any impact on the right to privacy associated with these measures. Overall, they improve the privacy of children whose personal data (such as name, photographs, location) would no longer be passively shared beyond users they are already connected with, or those that specifically search for the user. We therefore believe that these measures are compatible with the ICO Children's Code, as they are an age-appropriate application of common U2U service features and in that context, avoid using children's personal data in a way that might be detrimental to their safety and wellbeing.

18.69 We acknowledge that, to implement this mitigation, services would need to understand the age of their users and classify them as a child or adult. However, as outlined later in this chapter this measure does not require services to put age verification in place or obtain new personal data in this way.

## Who this measure applies to

18.70 Our analysis suggests that online grooming is a widespread and growing harm which can have a devastating impact on the lives of children. As we have explained, we believe that in general the measures in question would be effective in combating grooming and as a result bring about an important reduction in the sexual abuse of children. Whilst it is not possible to precisely quantify the benefits that would flow from this, our analysis suggests that they could be very significant. There may also be secondary benefits relating to the other kinds of illegal harms identified in our analysis and in particular harassment.

18.71 In relative terms, the direct monetary costs of the measures are likely to be fairly low for most services. The measures would also likely have a range of indirect costs that may have the effect of reducing use and, in turn, some services' revenues.

18.72 Given the significant impact grooming has and the fact that we deem the measures we are proposing would be effective in combating it, on balance, we consider that the measures are

likely to be proportionate when having regard to the expected impact on the risk of harm and the potential costs. We also consider that users would generally be likely to accept some of the inconveniences that could result from these measures if they were aware of the benefits that arise from the measures.

- 18.73 Although we broadly expect the measures to be proportionate, we recognise this may not be true for all services in all circumstances or there may be reasons why some services may not be able to practically implement the measures. Therefore, we have also considered which services should be applying these measures based on the extent to which:
- a) services have the relevant functionalities;
  - b) child users can be identified; and
  - c) the measures are proportionate for individual services.

### Services have the relevant functionalities


- 18.74 Some of the components of the measure depend on the services having particular functionality and therefore they would only need to apply that part of the measure if they have the functionality, e.g. the parts relating to setting defaults for network expansion and connection list functionalities would only need to be applied by services that had such functionality.

### Identifying child users

- 18.75 Clearly, the default safety setting measures, which would be applied only to child user accounts, would not be effective at reducing harm unless services know which of their users are children. It is therefore important to consider if and how services determine the age of their users. Our understanding is that the approach to determining user age varies significantly across the industry.
- 18.76 Certain services may not establish the age of their users at all. In this case, we could consider whether it would be proportionate to require the defaults to be applied to all users to effectively address the risk of considerable harm to children caused by online grooming, in circumstances where the alternative would be that the measures are not applied at all. While applying the measures to all users may secure the greatest effectiveness in terms of harm reduction, we are concerned that this would be ill-targeted and result in unintended consequences for both the service and its users.
- a) If the measure was applied to all users, the indirect costs<sup>534</sup> are likely to significantly increase as it would introduce friction and costs to all users on a service. This could lead to a reduction in the use of online platforms and cause a corresponding reduction in income from the services themselves. The benefits are also likely to increase at a much slower rate than the costs, compared to when the measures are more directly targeted at children, as any costs that are incurred by adults would not have any corresponding benefits in reducing the risk of grooming. The rationale for including this measure for all users is therefore much weaker if children cannot be identified.
  - b) While the interference with the right to freedom of expression and association of child users may be justified in view of the aims (outlined further in paragraph 18.66 to 18.68), we do not consider such intrusion to be justified or proportionate for adult users.

---

<sup>534</sup> As described above in paragraphs 18.52 to 18.65

- 18.77 In the circumstances, we provisionally consider that the potential impact on adult users outweigh the benefits of applying the measures to all users on service and that on balance, it would not be proportionate to expect services to apply these measures to all users if they do not have a means of identifying which of their users are children.
- 18.78 We therefore provisionally consider that services should only be in scope of this measure if they have existing means of identifying child users, whether that is a form of age assurance or another method. Services that do estimate whether users are likely to be children use a range of tools (subject to applicable data protection and privacy laws), these include:
- a) **Facial biometric age estimation** – where a user’s face is analysed and an age estimation is based upon the user’s features.
  - b) **Behavioural analysis** – algorithmic analysis of a user’s behaviour to estimate age.
  - c) **Age verification using hard identifiers** – This can include asking a user to input credit card details, open banking or capturing information from a photo-ID document uploaded by the user.
  - d) **Self-declaration** – a user declares their age and the service uses this as its basis for judging how old the user is.
- 18.79 Our understanding is that, at present, many user-to-user services that collect age information rely on self-declaration, which can be easily evaded through deliberate false declaration. Our research found that a third of respondents aged 8-17 who had a social media profile were pretending to be aged 18 or over<sup>535</sup>, suggesting that where a service relies on self-declaration to determine a user’s age, our recommendations would not be implemented for all children’s accounts. Perpetrators have also been shown to create accounts pretending to be children to groom and manipulate child victims.<sup>536</sup>
- 18.80 Nevertheless, our considered measures would still be effective to a significant extent because: 
- a) our research also indicated that two thirds of respondents accurately declared themselves to be under 18, suggesting there would still be significant potential benefits; and
  - b) With regard to false declarations by perpetrators, we have designed the measures under consideration with this risk in mind. For example, we are proposing that both adults and children should be blocked from sending direct messages to child users where their accounts are not connected.
- 18.81 We expect to consult on proposals relating to age assurance as part of our future phases of work, including in our consultation on the Protection of Children Code of Practice. **If robust age assurance measures were brought in for services, this would likely strengthen the effectiveness of these measures.**

---

<sup>535</sup> Ofcom, 2022. [A third of children have false social media age of 18+](#). [accessed September 4 2023].

<sup>536</sup> 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. [accessed September 21 2023].

## Proportionality of measures for different services – options

- 18.82 There remains a question about whether these measures are proportionate for all services or whether they should apply to a subset of services. This is complex. We have a relatively high degree of confidence that it would be proportionate to recommend the measures in question for the largest, riskiest services. However, the case becomes somewhat less clear cut the smaller the service becomes, and when it does not have the highest risk characteristics.
- 18.83 This is because, all else being equal:
- a) the fewer children there are on a service the lower the likelihood of grooming occurring on that service and so the potential benefits that arise from reducing the risk of grooming would also be commensurably lower;
  - b) the lower the risk characteristics of a service, the lower the chances of grooming occurring on that service; and
  - c) the smaller the service, the more difficult it is likely to be for the service to bear the direct costs of the measures and so there is a higher chance of a material negative financial impact on that service.
- 18.84 From paragraph A5.81 of our draft Risk Assessment Guidance in Annex 5, we explain that we consider that services with the following features are ordinarily likely to pose a high risk of grooming:
- a) Your service is likely to be high risk of grooming if it can be accessed by children **and** users are able to communicate one-to-one with child users (e.g. direct messaging);
  - b) **And any** of the following applies:
    - Your service has been systematically used by offenders for the purposes of grooming children for child sexual abuse;
    - Your service has a majority of risk factors associated with grooming in Ofcom’s U2U Risk Profile, in addition to child users and direct messaging.
    - Your service includes child users when users are prompted to expand their networks, including through network recommender systems (e.g. network expansion prompts);
    - Your service allows users to view child users in the lists of other users’ connections;
    - Your service has user profiles or user groups which may allow other users to determine whether an individual user is likely to be a child.
- 18.85 In addition, our Risk Assessment Guidance sets out when services are likely to identify as having a medium risk of grooming:
- a) Your service is likely to be **medium risk** of grooming if your service does not meet the criteria for high risk, **and**:
    - It can be accessed by children; and
    - users are able to communicate one-on-one with each a child (e.g. direct messaging).
  - b) **And any** of the following applies:

- Your service has recently been used by offenders for the purposes of grooming children for sexual abuse;
- Your service has two or more of the other risk factors associated with grooming in Ofcom's U2U Risk Profile, in addition to child users and direct messaging.

18.86 We have set out the expected riskiness of services in this way because, as set out above, publicly displaying children's details in this way allows perpetrators to more easily identify children to target and increase the likelihood that they are able to initiate the grooming process. In essence, the greater amount of information that is available about children online, the more likely it is that grooming will take place on a service.

18.87 We have taken account of the potential impact of service size and riskiness and considered the following options for how we could target the measures:

- Option 1:** Apply the measures to all large services which have a high or medium risk of grooming.
- Option 2:** Apply the measures to: (i) all services which have a high risk of grooming AND at least 25,000 child users; and (ii) all large services which have a medium risk of grooming.
- Option 3:** Apply the measures to: (i) all services which have a high risk of grooming and (ii) all large services which have a medium risk of grooming.

### Assessment of the options

18.88 We summarise our assessment of the options below, with more detail being given from paragraph A14.31 of Annex 14.

18.89 We are confident that Option 1 would be a proportionate intervention and that it would play an important role in combatting grooming on the largest platforms. As we have explained in the CSEA Register of Risks Volume 2: Chapter 6C (CSEA), paragraphs 6C7.29 to 6C7.30, one tactic deployed by some perpetrators is to target large services because significant numbers of children use them. This being the case and given the volume of children using these services and the prevalence of the harm, our analysis suggests that applying the proposed measures to large services which are high or medium risk for grooming<sup>537</sup> could result in significant reductions in grooming, thereby delivering material benefits.

18.90 As set out in more detail in Annex 14, estimating and quantifying the economic and social cost of CSEA offences, and the likely benefits of reducing grooming, is challenging. Our analysis indicates that the benefits are likely to be much greater than costs for large services with above 1 million child users, even though our estimate of benefits only considers the benefits that would arise from a reduction in contact CSA that is a result of online grooming. As there are other benefits, the total benefits from the measure for a service of this size are likely to be significantly higher.

18.91 We also assumed in the analysis that the expected costs were at the very top of our estimated range. This conservative approach combined with the high ratio of benefits to costs gives us significant confidence that applying the measures to all large services which have a high risk of grooming is likely to deliver significant benefits, and that Option 1 would

---

<sup>537</sup> We expect such large services would have more than 1 million child users. For an explanation, please see Annex 14, paragraph A14.62.



be proportionate. Additionally, we also considered our estimates in Annex 14 in the case of whether to apply the measure to large medium risk services. Although the analysis does not specifically cover medium risk services, our analysis for large high risk services show very high estimated benefits in comparison with cost. This gives us confidence that it is also likely to be proportionate for medium risk services, even though the benefit from the measure is likely to be slightly lower for medium risk services compared to high risk services.<sup>538</sup>

- 18.92 We have also considered the potential impact of the indirect costs outlined above and recognise that these costs are likely to be greater for large services, because they are likely to scale with the number of users on the platform. Despite this, we consider the measures are likely to be proportionate for large platforms as the potential magnitude of the benefits is so great relative to the direct costs we have quantified.
- 18.93 However, our provisional view is that Option 1 would not go far enough. There are a number of reasons for this:
- a) As set out in the Register of Risks, grooming is not confined to the largest platforms. Perpetrators can target any platform where there are children regardless of size.<sup>539</sup> Option 1 would therefore leave a material part of the problem of grooming unaddressed.
  - b) Moreover, it would likely have displacement effects. If the largest services take steps to improve protections against grooming, this may result in perpetrators shifting to focus on targeting children on smaller services. As we show in the Register of Risks, we have observed displacement effects of this nature occur when large services have moved to improve protections against other harms.<sup>540</sup>
- 18.94 The choice between Options 2 and 3 is more evenly balanced. Both options would be more effective at combatting grooming than Option 1. Option 2 would not apply the measure to services with very few children on them. This would reduce the risk of inadvertently imposing disproportionate costs on services where grooming did not in practice occur frequently. Option 3 would provide more comprehensive coverage against grooming, on all services which have identified as having a high risk of grooming and reduces the potential for displacement of perpetrators to very small services.<sup>541</sup>
- 18.95 Table A14.10 in Annex 14 shows that the estimated benefits are greater than costs across all four scenarios we have assessed for services with 25,000 child users.<sup>542</sup> One scenario (low benefit/high cost) shows benefits to be only moderately higher than costs, however as the benefits in the analysis are likely to be significantly understated,<sup>543</sup> we consider these results

---

<sup>538</sup> We expect the benefit to be slightly lower for medium risk services because we expect that they would have a slightly lower prevalence of grooming, that leads to contact CSA, than we have used in the analysis.

<sup>539</sup> Please see, Volume 2: Chapter 6(CSEA) paragraphs 6C.34 to 6C45.

<sup>540</sup> Volume 2: Chapter 6C (CSEA) paragraphs 6C7.21 to 6C7.23

<sup>541</sup> Note that this displacement could be either child users or perpetrators moving to smaller platform to get around restrictions. If large numbers of child users move to a smaller platform, we would expect it to come within scope of our measures as it grew in size, albeit with some lag. However, when perpetrators move to a smaller platform, no such 'correcting' effect will take place.

<sup>542</sup> All services captured under Option 2 would have at least 25,000 users.

<sup>543</sup> This is because we only consider the benefits that would arise from a reduction in contact CSA that is a result of online grooming and because our measure of estimated harm of individual contact CSA offences resulting from grooming is also likely to be understated for the reasons described in Annex 14.

indicate that, at the very least, the measure is likely to be proportionate for services with 25,000 child users or more.

18.96 Overall, we consider that the results of the quantitative analysis illustrate that the measures are likely to be proportionate for Option 2. However, what the quantitative analysis does not do is indicate whether the measures are also likely to be beneficial for services which have even lower numbers of child users, as it is unable to capture all of the factors that impact whether the measure is proportionate for those smaller services.

18.97 On balance, our proposal is to adopt Option 3. This has been informed by both the quantitative analysis and a wider qualitative assessment of the factors that affect the proportionality of the measure. The reasons for applying the measure to all high risk services are:

- a) The widespread nature of the threat grooming poses and the severity of the harms it can lead to. Child sexual abuse is a horrific crime which can have a severe and lifelong impact. This argues for applying Option 3 rather than Option 2, not least given that the impact of grooming is so material that the measure would only need to prevent a very small number of cases of grooming on any given service for the benefits to justify the costs of the measure.
- b) As described above, it is likely that the true benefit would be higher than we have been able to estimate, and potentially significantly higher than we have modelled for a service with 25,000 child users. This indicates it would be proportionate to apply the measure to services with potentially much lower numbers of child users.
- c) As set out in the Register of Risks, perpetrators target services of all sizes where there are children, even very small services.<sup>544</sup> Option 2 could therefore still leave important gaps in protection. A particular risk is the potential for perpetrators to move to using smaller services if it were easier to connect with children on such services because they were excluded from the measure. This risk is not captured by our quantitative analysis above.
- d) Option 3 only targets the measure at smaller platforms if they are at high risk of grooming. Given the severity of the harm, where a service is genuinely high risk there is a strong argument that it should not be exempt from providing children with protection, regardless of its size. The fact that the option places the most onerous obligations on the highest risk services is an important factor in our assessment of proportionality.
- e) Relatedly, we also consider that some small services that are high risk could be particularly unsafe to child users on that platform (ie, they have a higher risk of grooming than the average service we have included within our analysis). For those services, the proportion of child users who are targeted by perpetrators may be higher than the cross-sector averages we have calculated above. Consequently, the potential benefits (per child user) from introducing the measures are likely to be higher than is implied from the analysis. This further indicates that applying the measure to all services is likely to be appropriate to ensure we capture these types of services within the measures.
- f) The costs will tend to be towards the lower end of the range we estimated for smaller businesses because such businesses will not have material high overheads and

---

<sup>544</sup> Please see, Volume 2: Chapter 6(CSEA) paragraphs 6C.34 to 6C45.

coordination costs associated with implementing the measure. This strengthens the argument that our proposal is proportionate. For services which are not large (ie, those with a total user reach of less than 7 million) to be captured by our measures, they would need to both be able to identify child users and have assessed themselves as 'High risk' for grooming in their risk assessment.<sup>545</sup>

- g) It is likely that smaller services, and particularly if they are services with more risky network expansion and connection functions, are likely to be at a relatively early stage of development and in the process of growing their user base. We consider that there could be some benefit from providing certainty on the required safety measures when high risk functionalities are incorporated into online services, whatever the number of child users. This approach could be beneficial if it means services are not required to update systems and interfaces once they pass a certain number of child users. We also consider that for a new, growing service, the additional cost that it incurs from making a greater upfront investment due to the measures, compared to the costs which it would incur once it passes a certain size, is likely to be small.

## Provisional conclusion

18.98 Our provisional view is to recommend as a part of our CSEA Code and Other duties Code for U2U services that all services which identify a high risk of grooming and all large services that identify as at least medium risk of grooming in their risk assessment should ensure that, where relevant functionalities exist and they can identify child users, their default settings are such that:

- a) Children using a service are not presented with network expansion prompts, or included in network expansion prompts presented to other users.
- b) Children using a service should not be included in the connection lists of other users, and the connection lists of child users should also not be displayed to other users.
- c) For services with a user connection functionality, default settings should be implemented related to direct messaging, so that child users cannot receive direct messages from a non-connected user.
- d) For services with no user connection functionality, services should implement default settings so that child users should have 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 on the service, child users may be presented with an alternate means of actively confirming before any interaction associated with that functionality begins.
- e) The service should implement default settings which switch off automated location information displays for child users.

---

<sup>545</sup> Assessing as high risk for grooming includes having direct messaging functionalities alongside at least one other high risk factor, like network expansion prompts, connection lists or evidence of existing systematic grooming. Please further information on this please see Chapter 18, paragraph 18.85.

## Support for child users

---

### Harms that the measure seeks to address

18.99 As we have outlined in our assessment of the default safety settings, we consider there to be residual risks to children both as a result of the “default” nature of the settings, the ability of perpetrators to circumvent those measures, and the online communications that would not be restricted by those measures.

18.100 In particular:

- a) **A child user may seek to disable one of the recommended default settings:** We consider there to be residual risks associated with the default safety settings proposals for the measures set out previously in this chapter in respect of network expansion prompts, direct communication and location functionalities. As outlined above, our recommendations are choice preserving, meaning 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 measures are designed to address. Child users may choose to disable a default for many positive reasons, including finding new connections or followers; increasing reach of content they create; or having more features or functionalities available to them, among other things. Children may also wish to set their functionality settings to emulate adults. However, they could also experience pressure to turn off safety defaults by others on the platform, including potentially by adults engaging in grooming. We are therefore concerned about instances of child users seeking to disable these default settings without fully understanding the relevant grooming and other kinds of illegal harm risks they could be introducing.
- b) **A child would still have a choice to accept or deny a request from another user to establish a formal connection:** Children would 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 poses a particular risk as it can also increase the likelihood of children accepting and connecting with potential perpetrators, as outlined in our Register of Risk Volume 2: Chapter 6C (CSEA) paragraphs 6C7.47 to 6C7.49.<sup>546</sup> We recognise that by placing restrictions on children being able to receive direct messages from unconnected accounts it could cause an unintended rise in perpetrators seeking to establish formal connections with children as a means of circumventing the restriction.
- c) **A child user may engage in direct messaging with users, once they are connected:** Children would likely 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 our Register of Risk Volume 2: Chapter 6C (CSEA) paragraphs 6C7.59 to 6C7.64. As explained above, direct messaging can be exploited by perpetrators as a means of initiating the grooming process through

---

<sup>546</sup> For example: Anthony, R., Young, H., Hewitt, G., Sloan, L., Moore, G., Murphy, S. and Cook, S., 2022. [Young people’s online communication and its association with mental well-being. Results from the 2019 student health and well-being survey.](#) *Child and Adolescent Mental Health*, 28 (1). [accessed September 5 2023]; Fish J.N., McInroy L.B., Pacey M.S., Williams N.D., Henderson S., Levine D.S. and Edsall R.N., 2020. ["I'm Kinda Stuck at Home With Unsupportive Parents Right Now": LGBTQ Youths' Experiences With COVID-19 and the Importance of Online Support.](#) *J Adolesc Health*, 67 (3). [accessed September 5 2023].

private communications. Although, children may be aware of risks and harms when interacting with people online, they are often unsure how to avoid them.<sup>547</sup>

- d) **A child user may not feel able to end contact with users when they feel unsafe or uncomfortable.** Research suggests that children find it difficult to end contact with perpetrators of grooming or other CSEA offences online; sometimes, as a result of blackmailing or threats they are experiencing.<sup>548</sup> This is despite children being more likely to use online reporting tools compared to turning to offline support systems such as a caregiver or a friend.<sup>549</sup> Research also indicates that children are more likely to block users than report them as they feel unclear of the process, which can discourage future reporting.<sup>550</sup> When trying to end contact with a perpetrator, victims of grooming often recall feelings of anxiety, fear or worry that their images will be distributed online, or that the offender will try to recontact them.<sup>551</sup> It is often the case that these threats and coercive tactics have significant impact on the victim's mental health, which can include self-blame, negative sense of self, depression, anxiety, and suicidal ideation. There have been recent reports of victims taking their own lives following incidences of grooming and perpetrators threatening to share their sexual images.<sup>552</sup> At the point of taking 'action,' child users are likely to feel a heightened sense of vulnerability if perpetrators have used threats.

## Options

- 18.101 To increase the efficacy of the safety default measures and help to reduce these residual risks, we have considered options for providing children with timely and accessible information at critical points in their user journey to enable them to make informed choices about risk in their online experiences.
- 18.102 We have identified four critical points where information could be presented to child users to increase their understanding of the risks. These are:
- When a child user seeks to disable one of the recommended default settings;** The information provided should assist children in understanding the implications of making this change, including the protections afforded by the default setting they are disabling.
  - At the point where a child is making a choice to accept or deny a request from another user to establish a formal connection;** The information provided should explain the

---

<sup>547</sup> 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 September 4 2023].

<sup>548</sup> 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 September 4 2023].

<sup>549</sup> Thorn, 2021. [Responding to Online Threats: perspectives on Disclosing, Reporting, and Blocking](#). [accessed September 4 2023].

<sup>550</sup> *ibid.*

<sup>551</sup> Joleby M, Lunde C, Landström S and Jonsson LS, 2020. ["All of Me Is Completely Different": Experiences and Consequences Among Victims of Technology-Assisted Child Sexual Abuse](#). *Frontiers in Psychology*, 11. [accessed September 21 2023].

<sup>552</sup> Dearden, L, 2018. [Five British men have killed themselves after falling victim to online 'sextortion', police reveal](#). *The Independent*, 14 May. [accessed September 4 2023].

types of interaction that would be enabled through establishing a connection, and how to take action against a user.

- c) **At the point where 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.
- d) **At the point where a child user is taking action against another account, including blocking, muting or reporting;** 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.

18.103 We provide more detail on the information on the proposals, including on the information that should be provided, when discussing efficacy below.

## Effectiveness

### Effectiveness of providing information at relevant times

18.104 As with defaults, a prompt can be used to influence user behaviour, in this case to improve their safety, at the same time preserving the option for the user to choose another course of action. We consider that the provision of information of our option would help children make more informed choices regarding their safety by giving them the information they need at the right time.

18.105 Academic and regulatory work suggests that prompts can influence people to make safer choices.<sup>553</sup> Many respondents to our Call for Evidence flagged the existence and potential of additional ‘prompts’ that can be served to users as they navigate online services.<sup>554</sup> There is some evidence that prompts can work for children in particular. Academic 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 online (tending to be ‘trusting, naïve, curious, adventuresome, and eager for attention and affection’).<sup>555</sup>

18.106 While some research indicates that prompts can be effective, other research suggests that they can be perceived as ‘annoying’,<sup>556</sup> and there are concerns that excessive frequency could lead to alert fatigue where people do not engage with the information.<sup>557</sup> However

---

<sup>553</sup> For example: European Commission, 2019. [Study on media literacy and online empowerment issues raised by algorithm-driven media services](#). [accessed September 21 2023]; US Food and Drug Administration, 2019. [Communicating Risks and Benefits: An Evidence-Based User's Guide](#). [accessed September 4 2023]; Tussyadiah I., Miller G., Li S. and Weick M., 2021. [Privacy nudges for disclosure of personal information: A systematic literature review and meta-analysis](#). *PLoS One*, 16 (8). [accessed September 21 2023]; Acquisti et al., 2017. [Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online](#). *ACM Computing Surveys*, 50 (3). [accessed September 4 2023].

<sup>554</sup> ICO; Meta; [CONFIDENTIAL X]; [CONFIDENTIAL X]; and Roblox responses to 2022 Illegal Harms Ofcom Call for Evidence.

<sup>555</sup> Alemany, J., del Val E., Alberola, J., García-Fornes, A., 2019. [Enhancing the privacy risk awareness of teenagers in online social networks through soft-paternalism mechanisms](#). *International Journal of Human-Computer Studies*. 129. [accessed September 4 2023].

<sup>556</sup> 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 29<sup>th</sup> Australian Conference on Computer-Human Interaction. [accessed 21 September 2023].

<sup>557</sup> ICO, 2021. [ICO to call on G7 countries to tackle cookie pop-ups challenge](#). [accessed September 4 2023].



despite users finding the frequency of the prompts bothersome, we do not consider that means they are ineffective in meeting the desired objectives.

- 18.107 There is substantial evidence that the timing and relevance of such interventions is particularly important to ensure that they achieve the desired effect.<sup>558</sup> This is consistent with what has been advocated by others:
- a) The Australian eSafety Commissioner, in their Safety by Design principles, 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.”<sup>559</sup>
  - b) 5Rights 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.<sup>560</sup>
  - c) In an article for TTC Labs, Dr Dan Hayden, data strategist at Meta, 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’.<sup>561</sup>
- 18.108 This suggests that, to optimise the benefits, it is important to consider the right time to serve a prompt.
- 18.109 In light of the evidence above, we consider that providing relevant information at certain critical points is likely to be effective at mitigating the risks identified in our discussion of harms and risk above and would contribute to the reduction of grooming.
- 18.110 Below we discuss in more detail why we consider that the provision of information at the critical points we have identified would be particularly effective at mitigating and managing the risk of harm to child users online.

### Information when disabling default settings

- 18.111 We believe that there are benefits to providing child users with information regarding the potential risk involved, at the point of disabling the safety default settings recommended earlier in this chapter. The information should assist child users in understanding the implications of disabling that default setting, including the protections it affords. This would help to ensure that child users understand the implications of making this change, including a reminder of the protections afforded by the default setting they are disabling.
- 18.112 Once informed of these risks, users would have an opportunity to change their minds about disabling the default safety settings. It would also 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 settings because of pressure or blackmail.
- 18.113 We are not aware of prompts currently being used comprehensively to warn children of the risks of disabling default safety settings. Nonetheless, in light of the evidence on the user

---

<sup>558</sup> The Behavioural Insights Team (Costa, E. and Halpern, D.), 2019. [The Behavioural Science of Online Harms and what to do about it](#). [accessed September 4 2023].

<sup>559</sup> Australian e-Safety Commissioner, 2019. [Safety By Design Principles and Background](#). Principle 2.3 [accessed September 5 2023].

<sup>560</sup> 5Rights, Alan Turing Institute and Glitch responses to Ofcom 2022 Illegal Harm Call for Evidence.

<sup>561</sup> TTC Labs (Hayden, D.), 2021. [Making Sense of Data Disclosures](#). [accessed September 5 2023].



safety benefits of prompts, we believe that providing support at this stage of the user journey would aid in reducing the risk of harm to child users.

### Information in early interactions with a user

18.114 We believe that presenting certain information to child users at the point of interactions that present a higher risk of grooming, would support child users to make informed choices about their engagement with the service and other users. We provisionally consider that there are two points in the user journey at which information as outlined below would be effective at achieving these aims:

a) **At the point where a child is making a choice to accept or deny a request from another user to establish a formal connection:** This would only apply to services that have formal user connections on their services (such as friends, followers, and connected users), where such connections are conditional on requests being accepted by another user.

18.115 We provisionally consider that the following information would be beneficial for child users at this point of the user journey:

- i) Information on the types of interactions that would be enabled through establishing a formal connection, including that the user will be able to communicate directly with them or view and engage with shared content, if this is applicable to the platform.
- ii) Information on how to take action against, or restrict interaction with, another user. This could include signposting to blocking and reporting tools.

b) **At the point where the first direct message is exchanged (either sent or received) between a child user and another user.**

18.116 We provisionally consider that the following information would be beneficial for child users at this point of the user journey:

- i) A reminder that this is the first time they are communicating with this user in a one-to-one environment on this platform.
- ii) Information on how to take action against, or restrict interaction with, another user. This could include signposting to blocking and reporting tools.

18.117 We anticipate that the informational prompts described above would provide friction during two critical, early points in the grooming journey in which perpetrators may seek to establish contact with a child user. We expect that the prompts, as well as providing the child user with information to support them to make informed choice about their contact on the platform, may cause the child user to pause before engaging with a new user. The provision of information about how to end contact with another user, such as through blocking or reporting, could help equip children with the knowledge of how to protect themselves should their future engagement with that user cause them to feel uncomfortable or unsafe.

18.118 As outlined in paragraph 18.35 we provisionally 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 that functionality begins.

## Information when seeking to take action against a user

- 18.119 As explained in paragraph 18.98 (d) above, ending contact during online grooming is often one of the most sensitive and distressing times for a child and, for a host of complex reasons, child users may be hesitant to do so. They often do not believe that the grooming experience is severe enough or important enough to report, they may experience embarrassment or shame, feel they will not be believed, and they may worry about the repercussions if a perpetrator finds out they have taken action against them.<sup>562</sup>
- 18.120 We therefore consider that the provision of information when a child user is seeking to take action to block, mute or report another user (as relevant to a service) may be beneficial to child users that have had a negative experience and require additional knowledge to make informed decisions to support their ongoing safety on the platform. This can also have positive ramifications for child users feeling safer offline after having taken action on the platform. The following could therefore support these aims:
- a) Information on the effect of the action taken on interactions with the user in question. For example, an explanation of the communication functionalities that would be restricted (such as comments and direct messaging) and where applicable, confirming whether the user will be made aware of the fact that the child has taken action against them.
  - b) Information on further steps the child user can take to limit interactions or increase their safety on the service. This could include information on how the child can review their privacy settings and information on other actions they can take against the user should they wish to.
- 18.121 Taking action against an account, whether that is reporting or blocking, can often increase the risk towards a child in a grooming scenario, for example the offender increasing or carrying out their threats toward a child.<sup>563</sup> It is anticipated that by providing child users with clear information on the effect of an ‘action’ on their interactions with the user in question, children may feel safer on the service as they will understand to what extent that user can communicate with them or access their information. They may also feel encouraged to proceed to take action if the service confirms that the user would not be made aware.
- 18.122 We also believe that it would be beneficial for children to be made aware of other options to manage risk at the point of wanting to take action against another user, as it would equip them with knowledge of the actions they can take to feel safer. For example, at the point of a child user blocking another user, the child may receive information on other platform settings which may empower the child user to increase privacy settings on their account. This would, in turn, help prevent the perpetrator being able to send them messages from other accounts.
- 18.123 This timely information may also encourage children to report a bad actor. We understand that most children will prefer to block rather than report a user, which may be, in part, because reporting can often feel like an extreme step to a child. The provision of information about options to take action against a user may help them to make a more informed choice about what action, if any, they want to take. It may then be that children feel more empowered to report a user, which offers benefits beyond just blocking or muting, as the

---

<sup>562</sup> Thorn, 2021.

<sup>563</sup> Katz, C., Piller, S., Glücklich, T. and Matty, D. E., 2021. [“Stop Waking the Dead”: Internet Child Sexual Abuse and Perspectives on Its Disclosure](#). *Journal of Interpersonal Violence*. 36 (9–10). [accessed September 4 2023].

service would be made aware of potential harmful actors. Effective reporting can help to notify a service of potential individual bad actors and serve as a signal that grooming behaviours are occurring on site potentially at a systematic level and informs services they may need to take actions they might otherwise not have realised were needed.

## Format of prompts

- 18.124 We recognise that another factor that may determine the effectiveness of the information provided is the format in which it is delivered. However, the available evidence regarding how to present user support messages does not point to a single ‘best practice’ approach.
- 18.125 In the many studies that found similar mitigations to be effective, the authors pointed to various factors that they considered to contribute to the effectiveness of using prompts, such as length, colour and language, but there is no consistent recommendation on *how* to apply these factors.<sup>564</sup> The ICO in its Children’s Code says providers “should bear in mind children’s needs and maturity will differ according to their age and development stage” and it provides a guide for considering the interests, needs and evolving capacity of children at different ages.<sup>565</sup>
- 18.126 There is also some variation in the technical interfaces through which services communicate safety prompts to users in their current practice. Some present this information as a pop-up, while others embed it within an interface. At this stage, we do not have sufficient evidence to understand any differences in effectiveness of these approaches.
- 18.127 Overall, given the lack of conclusive evidence, we provisionally consider that services are better placed than Ofcom to design, test, and evaluate the format and delivery of the prompts to optimise the benefits for child users. As such, we are not proposing to make specific recommendations around how the supportive knowledge should be presented and encourage services to establish their own best practice on how to deliver information to child users.
- 18.128 We do, however, expect that the information is easy for child users to understand and is displayed prominently to them at the relevant critical point. We also consider it proportionate to provide some guardrails as to the nature of the information that should be provided. This would ensure that the measure is effective at increasing the risk awareness of child users and ensuring that they can make informed, risk conscious decisions in their online interaction.

## Costs and risks

- 18.129 Direct costs on services are likely to be largely one-off costs. They would consist of developing information to present, and system changes to implement its appearance at one of the four critical points, such as before the disablement of the default is finalised. There would also be some on-going costs to maintain the functionalities. We are not proposing to prescribe precisely how services should provide information. As a result, 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. For some

---

<sup>564</sup> For example: Ioannou et al, 2021. This literature review called for further research “to elucidate the relative effectiveness of different intervention strategies and how nudges can confound one another”.

<sup>565</sup> ICO, 2022. [Age Appropriate Design Code](#). [accessed 21 September 2023]. We refer to this as the ‘Children’s Code’.

services, particularly larger, more sophisticated services, there may also be costs involved with testing and evaluating the format and delivery methods, with the possibility of changing these if they are not working well.

- 18.130 Software development costs could be material if services do not already have a system to provide prompts. They could, however, be much lower if they have an existing system to provide warnings/interstitials in other contexts.
- 18.131 We assume that the software development of applying this measure would take up to approximately six to 36 months of staff resources, made up of both software engineering time and other professionals (eg, project management). The exact cost would depend on the complexity and existing functions of the system and the extent of the supportive information that is provided. We expect this to cost in the region of £25,000 - £300,000.<sup>566</sup>
- 18.132 For smaller services, we would expect the costs would tend to be towards the lower end of the range. This is because smaller business will tend to have lower overhead and coordination costs in making changes.
- 18.133 As above, we assume the annual running costs of this measure to be 25% of the initial implementation costs. Therefore, the annual running costs would be approximately £6,250-£75,000. Ongoing running costs are likely to include regular updating of the supportive information and miscellaneous system maintenance costs. We recognise that these safety interventions may result in costs to users in terms of additional time and effort. However, in general we expect these user costs are to be relatively small for these intervention as these information points for children not designed to be frequent or overly intrusive.
- 18.134 It is plausible that there could be wider costs for services if the measure resulted in a reduction in the use of a service. It could alter the flow of the user experience for child users which could have an impact on engagement and therefore impose some indirect costs on the service, but that our provisional view is that this is likely to be proportionate given we think it will materially improve child safety online.

## Rights impacts

### Freedom of expression and freedom of association

- 18.135 We recognise that these user support measures may have a limited chilling effect on the rights to freedom of expression and freedom of association in that they would briefly delay children from disabling defaults and may result in children being less likely to do so (preserving the existing restrictions on their rights outlined in paragraph 18.65 above). The measures may also result in children being less likely to establish new connections or communicate with new users online. However, we expect the delay to the child concerned would generally be negligible and they would not be prevented from disabling the default settings if they so wish, nor would they be prevented from adding or communicating with new connections. The restriction on the ability of adults to impart information and ideas to the child concerned would be an informed choice of that child, and not something that we propose to require through these measures.
- 18.136 As such, we consider it unlikely that this amounts to an interference with the rights to freedom of expression or association. If it did, it would be proportionate to the overall

---

<sup>566</sup> This is based on the assumptions, such as for salaries, set out in Annex 14, paragraph 14.5.

reduction in the risk of harm that is achieved by increasing children’s awareness of the risk associated with certain activities on a service.

## Privacy

18.137 We do not anticipate that this mitigation would have an impact on the right to privacy, as we have not recommended as part of this measure that services extract or retain information relating to an individual’s engagement with, or action following, the provision of information beyond that which they would have done in the normal running of the service and the action of changing a setting would happen anyway.

## Provisional conclusion

18.138 Our analysis above illustrates the significant benefits that can arise from providing information to child users at critical points during their use of the service. Broadly speaking there are three aspects to the benefits that arise from the provision of this information:

- a) **Information leading to a reassessment of a user choice** – For example, a child may decide not to turn off a safety default after being provided with information that informs them of the potential risks involved.
- b) **Information leading to more awareness and knowledge of the potential risks when interacting with other users online** – For example, a child may accept a message from an unknown user, but with greater awareness of how to take action against a user on a service if they feel uncomfortable.
- c) **Information that leads to a child feeling safer online** – For example, individual child users are likely to become more informed, empowered, and supported with the information provided.

18.139 We consider that increasing the information available to children in order to realise these benefits is a particularly important tool to reduce the risk of grooming. We are confident that it would be proportionate to require the largest platforms on which grooming can occur to provide these messages, as the potential benefits would be significant given the large numbers of children on those services.

18.140 As with the default setting measure, we have considered whether to apply to all high risk U2U services or whether to exclude the smallest services by using a threshold based on the number of child users. On balance, we consider it is proportionate to apply these measures to all services which have identified a high risk of grooming as part of their risk assessment. The reasons for this are similar to that for the default setting measure above:

- a) As set out in the Register of Risks, perpetrators can target services of all sizes where there are children, even very small services.
- b) The widespread nature of the threat grooming poses and the severity of the harms it can lead to. Child sexual abuse is a horrific crime which can have a severe and lifelong impact. This argues for applying more widely, not least given that the impact of grooming is so material that the measure would only need to prevent a very small number of cases of grooming on any given service for the benefits to justify the costs of the measure.
- c) A particular risk is the potential for perpetrators to move to using smaller services if it were easier to connect with children on such services because they were excluded from the measure.

- d) We are only proposing to recommend smaller platforms to implement this measure if they are high risk of grooming. Given the severity of the harm, where a service is genuinely high risk there is a strong argument that it should not be exempt from providing children with protection regardless of its size. The fact that the option places the most onerous obligations on the highest risk services means that we consider this would be proportionate.
- e) The costs will tend to be towards the lower end of the range we estimated for smaller business, because such business will not have such high overhead and coordination costs. Moreover, services would have a significant degree of flexibility in the type of information and how it is provided, allowing services to develop an approach that is appropriate for their circumstances. This tends to mitigate the impact on services' costs. Although we propose to recommend the provision of this information, we are not recommending that users should have to confirm that they have read or seen this information. This means it is likely to be less intrusive to users and reduces the disruption to the user's experience on the service.

18.141 We also consider that this would be proportionate for large services (ie, services with over 7 million monthly UK users) that have identified as having a medium risk of harm. These services are likely to have a much greater number of child users and so the provision of information is likely to provide benefits to a wider number of child users. These services are also likely to have greater capacity to implement these measures.

18.142 Our proposed approach is therefore consistent with our approach for the default setting measures outlined above and ensures that all services that have the highest risk of grooming would be expected to provide information at critical points in the user journey.

18.143 We therefore propose to recommend as a part of our CSEA Code for U2U services that all U2U services that identify as high risk of grooming, or large services that identify as medium risk of grooming, provide information to child users to assist them to make informed choices about risk and information to access safeguarding support at four critical points on their online journey –

- a) When a child user seeks to disable one of the recommended default settings, services should provide child users with information that assists them in understanding the implications of disabling a that default setting, including the protections afforded by the default setting they are disabling.
- b) At the point where a child is seeking to respond to a request from another user to establish a formal connection, services should provide the following information to child users before the connection is finalised:
  - i) the types of interactions that would be enabled through establishing a connection; and
  - ii) the options available to take action against a user, such as blocking, muting, reporting or equivalent action.
- c) At the point where a child user engages in direct messaging with another user for the first time, a service should provide a child user with the following information:
  - iii) a reminder that this is the first direct communication with that user; and
  - iv) the options available to take action against a **user**, such as blocking, muting, reporting or equivalent action, unless direct messaging is a necessary and time critical element of another functionality of the service, in which case a child user

may be provided this information before any interaction associated with that functionality begins.

- d) At the point a child user is taking action against another account, including blocking and reporting, a service should provide a child user with the following information:
  - v) the effect of the action, including the types of interactions that it would restrict and whether the user would be notified; and
  - vi) the further options available to limit interaction with the **user** or increase their safety.

This information should be prominently displayed and be clear and easy for a child user to understand.



# 19. Recommender system testing (U2U)

## What is this chapter about?

Recommender systems are a primary means through which user-generated content is disseminated across U2U services, and the means via which users encounter content. This chapter discusses steps U2U services can take to monitor and manage the illegal content risk posed by their recommender systems.

## What are we proposing?

When services make changes to their recommender systems, they often carry out on-platform tests to assess the impact those changes will have. We understand that these tests typically focus on the impact design changes will have on commercial and engagement metrics.

We are making the following proposals for U2U services which already carry out on-platform tests of their recommender systems and that identify as medium or high risk for at least two specified harms<sup>567</sup>:

- **Services should, when they undertake on-platform tests, collect safety metrics that will allow them to assess whether the changes are likely to increase user exposure to illegal content.**

## Why are we proposing this?

Recommender systems can be found on many types of U2U service and are often essential to ensuring that users encounter content they enjoy and are likely to engage with. However, where illegal content is uploaded to a U2U service and missed by any content moderation systems that are used at the point of upload, recommender systems may play a role in amplifying the reach of that illegal content and increasing the number of people who encounter it.

In our Register of Risks, we identify that the way in which recommender systems are designed can influence the extent to which illegal content is disseminated on a service.

Gathering information about the impact changes to recommender systems have on the dissemination of illegal content will put services in a position to make materially better design choices than they otherwise would. This should reduce the online harm users experience.

Given that we are focusing this measure on services that already conduct on-platform tests, our provisional view is that services in scope of the measure are likely to be able to absorb these costs relatively easily. Whilst this measure may impose some costs on services, it may also deliver some countervailing savings as identifying and addressing potential causes of harm upfront may reduce the costs services incur mitigating harm after the fact. For example, reducing the extent to which recommender algorithms disseminate illegal content may reduce the costs content moderation teams incur dealing with reports of illegal content.

## What input do we want from stakeholders?

---

<sup>567</sup> CSAM; extreme pornography; intimate image abuse; foreign interference; terrorism; encouraging or assisting suicide or serious self-harm; hate; harassment, stalking, threats and abuse.

- Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.
- What evaluation methods might be suitable for smaller services that do not have the capacity to perform on-platform testing?
- We are aware of design features and parameters that can be used in recommender system to minimise the distribution of illegal content, e.g. ensuring content/network balance and low/neutral weightings on content labelled as sensitive. Are you aware of any other design parameters and choices that are proven to improve user safety?

## Introduction

19.1 In meeting their illegal content safety duties, U2U services are required to consider measures that relate to the “design of functionalities, algorithms and other features”. In this chapter we focus on steps that services can take regarding the design of their recommender systems, which are underpinned by algorithms.

### Box A: What is a recommender system?

By **recommender system**, we mean a system that determines the relative ranking of content on a U2U service. The measure considered in this chapter would only apply to recommender systems that are used for the curation of user-generated content feeds, for example newsfeeds and reels on certain services. These are known as content recommender systems. The measure would not apply to recommender systems that underpin search functionalities on a U2U service, or network recommender systems that suggest other users to follow or groups to join.

19.2 Recommender systems are deployed across many types of U2U service and are often essential to ensuring that users encounter content they enjoy and are likely to engage with.. However, where illegal content is uploaded or shared to a U2U service and missed by any content moderation procedures that are engaged at the point of upload, recommender systems may play a role in amplifying the reach of that illegal content and increasing the number of people who encounter it.<sup>568</sup>

19.3 Across our Register of Risks chapters<sup>569</sup>, we identify that the way in which content recommender systems are designed can influence the extent to which certain categories of illegal content are disseminated on a service.<sup>570</sup> As such, design changes to recommender systems can also influence how much illegal content is served to end users. Box B below, sets out a number of examples of design changes that could in theory have that effect.

<sup>568</sup> Ofcom, 2023. [Evaluating recommender systems in relation to illegal and harmful content](#). [accessed 26 September 2023]. Subsequent references throughout.

<sup>569</sup> See Volume 2, Chapter 6 – Part 1.

<sup>570</sup> We also discuss this below. See paragraph 19.7 for further detail.

### Box B: What do we mean by design change?

By **design change**, we mean an alteration that is made either to the recommender system’s underlying model(s) or to the pool of content analysed and processed by those models. A design change could include, but is not limited to:

- **Expanding/constraining the content pool:** This means altering the system so that it changes the range of content it processes and recommends to users (e.g., altering a system so that it analyses content from all accounts, not just those a user follows). This is also known as the sourcing criteria.
- **Adjusting content signals:** This involves changing the types of cues the system considers when ranking content (e.g., altering a system so that it analyses how much a piece of content has been liked, viewed, and reposted, to determine its relative ranking).
- **Tuning prediction weights:** This refers to changing the emphasis the system places on different predictions made by the model (e.g., so that the system places greater weight on how likely a user is to comment on a piece of content, versus other predictions such as how likely a user is to share that content).

- 19.4 Research commissioned by Ofcom indicates that it is common for U2U services to make frequent alterations to the design of their recommender systems, with some services making hundreds of design changes every week.<sup>571</sup> Our concern is that some services may implement these changes without fully considering the illegal content risk to users because the risk assessment duty in the Act is only triggered by ‘significant’ changes.<sup>572</sup>
- 19.5 Ofcom commissioned research<sup>573</sup> has suggested a variety of measures that could be taken to manage this risk. Among them are: improving the transparency of how recommender systems operate (e.g., by giving users plain English explanations of why they are being served particular types of content); establishing more robust record-keeping procedures so that service engineers better understand how their own systems operate at a technical level; and ensuring that the “goal criteria” of recommender systems is not solely optimised to increase engagement (e.g., including objectives such as network balance, diversity, novelty, recency, and serendipity). While these measures may have merit, we do not believe we have sufficient evidence at this stage to justify proposing their inclusion in the draft illegal content Codes of Practice.
- 19.6 We do, however, believe that on-platform testing (see Box C below) can be an effective means of minimising the risk of user exposure to illegal content. Many U2U services, particularly the largest ones, already conduct on-platform tests to understand the likely consequences of making a change to the design of their recommender system, with a focus on understanding the effects on commercial and engagement metrics (e.g., on the number of likes, shares, advert clicks, and time spent on service). Given limited evidence on the

---

<sup>571</sup> Ofcom, 2023. [Evaluating recommender systems in relation to illegal and harmful content](#). [accessed 26 September 2023].

<sup>572</sup> Ofcom 2023. [Illegal content risk assessment guidance](#). [accessed 26 September 2023].

<sup>573</sup> Ofcom, 2023. [Evaluating recommender systems in relation to illegal and harmful content](#). [accessed 26 September 2023].

efficacy of the evaluation methods mentioned above, we have focussed on on-platform testing.

### Box C: What is on-platform testing?

By **on-platform testing**, we mean the process of testing two or more variants of a recommender system before proceeding with a design change. During testing, services collect data that can then be used to produce metrics. On-platform tests are set up and executed in a testing environment. On-platform testing methods include (but are not limited to):

- **A/B/x Testing:** This is a randomised control trial where the service creates a treatment group of users who are served content from the altered recommender system(s), and a control group of users who continue to be served content from the current recommender system. The results are then compared, and a decision is taken whether to implement the new variant.
- **Multi Arm Bandit (MAB) Testing:** Unlike A/B/x tests, which have static control and treatment groups, MAB testing is a randomised control trial that uses machine learning techniques to allocate users to the “best” performing variant of a recommender system while the test is underway (e.g., the branch that generates the most engagement, or another meaningful performance metric).

## Monitoring safety metrics when testing recommender systems

---

### Harm that the measure seeks to address

- 19.7 Where illegal content is shared on a U2U service, its recommender systems (where used) may play a role in disseminating that content to users. The specific risk we are concerned with here, and one noted in our Register of Risks,<sup>574</sup> is that services make changes to the design of their recommender systems in a way that results in users being more likely to be exposed to illegal content.<sup>575</sup> Services may make these design changes without understanding the likely risk to end users of their service.
- 19.8 Several studies and journalistic reports highlight the role of recommender systems in the dissemination of illegal and other harmful content. A working group review from the Global Internet Forum to Counter Terrorism<sup>576</sup> highlighted that there is a consensus among experts in the technology, government, civil society, and academic sectors that supports claim. While the focus of these studies tends to be on harmful content, our view is that illegal content would be disseminated in a similar way, given how recommender systems rank and curate content. This evidence includes:
- a) A systematic review by Yesilada and Lewandowsky, which determined that the content recommender system on one platform can facilitate pathways towards radicalising and extremist material. The study was not able to attribute this risk to certain design

---

<sup>574</sup> Volume 2, Chapter 6 – Part 1.

<sup>575</sup> Ofcom, 2023. [Evaluating recommender systems in relation to illegal and harmful content](#). [accessed 26 September 2023].

<sup>576</sup> Global Internet Forum to Counter Terrorism 2021. Content-Sharing Algorithms, Processes, and Positive Interventions Working Group: Part 1. [accessed 27 September 2023]

characteristics due to limited researcher access, however it did highlight the need for improved transparency and auditing procedures that can uncover the design characteristics that facilitate user-pathways towards problematic content.<sup>577</sup>

- b) An investigation by journalists Cook and Murdock, which identified that platform users on one platform could be led on recommendation trails from soft-core pornography to content featuring partially clothed minors. The study found that there was a progression of recommendations from videos showing adult nudity, to those featuring minors in sexualised contexts.<sup>578</sup>
- c) A study from Whittaker et al. (2021), which used digital avatar accounts to examine how recommender systems impacted user exposure to extremist content on different platforms.<sup>579</sup> This found that while some recommender systems did disseminate extremist and so-called ‘fringe’ content, others did not, indicating the importance of different design choices on the risk of encountering harmful content.<sup>580</sup>

19.9 It is our understanding that recommender systems are more likely to disseminate illegal content that is posted publicly and can receive user engagement signals in the form of likes, shares, and comments. Our Register of Risks identifies that content recommender systems can increase the risk of certain types of illegal content appearing on a service.<sup>581</sup>

## Options

19.10 One way of managing and monitoring the risk of amplifying illegal content would be through establishing on-platform testing of recommender systems or expanding their scope if used already. This would enable services to improve their understanding of the likely consequences of their design changes, and in turn allow for more informed decision-making when deploying changes in future.

19.11 We have considered the following options:

- a) Option A: All U2U services should develop and carry out on-platform testing of their recommender systems. As part of this, services should:
  1. **Produce additional safety metrics** to understand which variant of a recommender system is more likely to disseminate illegal content (the proposed safety metrics are explained below in more detail).
  2. **Keep a log of the test results**, noting the performance of each variant of the recommender system across the safety metrics, a description of the change to the recommender system being tested, and an explanation of the decision that was taken at the end of the test. This should give a reasonable indication of which design changes contributed to an increase or decrease in the dissemination of illegal content.

---

<sup>577</sup> Yesilada, M., and Lewandowsky, S., 2022. [Systematic review: YouTube recommendations and problematic content](#) *Internet Policy Rev*, 11 (1). [accessed 27 September 2023].

<sup>578</sup> Cook, J. and Murdock, S., 2020. [YouTube is a Pedophile’s Paradise](#). *Huffington Post*, 20 March. [accessed 26 September 2023].

<sup>579</sup> Whittaker, J., Looney, S., Reed, A., Votta, F., 2021. [Recommender systems and the amplification of extremist content](#), *Internet Policy Review*, 10 (2). [accessed 26 September 2023].

<sup>580</sup> Different platforms are likely to have varying volumes of illegal content present at any given period, and this (as well as a design choices) can be an influencing factor in the extent to which their recommender system may disseminate such content.

<sup>581</sup> See Volume 2, Chapter 6 – Part 1.

3. **Consult the log before making future design changes:** 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 before making future design changes.
- b) **Option B: Restricting the recommendation in Option A to only those U2U services that already employ on-platform testing of their recommender systems.**
  - c) **Option C: Restricting the recommendation in Option A to U2U services that both (i) already employing on-platform testing of their recommender systems and (ii) have also assessed that they are high or medium risk in their latest illegal content risk assessment for at least two of the kinds of illegal harms identified in paragraph 19.53.**

## Outline of the measure

- 19.12 Based on our understanding of the available methods of evaluating recommender systems, we recommend that services produce the safety metrics set out in Table 19.1 (or equivalent) across control and treatment groups when running an on-platform test on their recommender systems, where those systems curate and serve content to UK users.
- 19.13 Our research indicates that the complaints and reach data required to produce these safety metrics is already collected and tracked by many U2U services.<sup>582</sup>

**Table 19.1: Safety metrics**<sup>583</sup>

Metric	Description
<b>Total number of content items identified as illegal content or as an illegal content proxy</b>	The total number of items of content that are identified as illegal content or as an appropriate illegal content proxy (defined in paragraphs 19.20 – 19.25 below) in response to a user complaint during testing.
<b>Total number of impressions and reach per item identified as illegal content or an illegal content proxy</b>	For each piece of content identified as illegal content or an illegal content proxy: <ul style="list-style-type: none"> <li>• Impressions: the total number of times that content was displayed to users.</li> <li>• Reach: the total number of unique users that the content was displayed to.</li> </ul>

Source: Ofcom analysis

- 19.14 To produce these metrics, services would need to measure the dissemination of illegal content that has been identified in response to user complaints made during the testing window.
- 19.15 We recognise that services 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. We therefore propose in Chapter 12 to recommend that it would be an appropriate action, in

<sup>582</sup> Ofcom, 2023. Ofcom, 2023. [Evaluating recommender systems in relation to illegal and harmful content](#). [accessed 26 September 2023].

<sup>583</sup> By metric we mean a descriptive statistic that is used to monitor and evaluate behaviours or other phenomenon across a population or environment. Metrics are typically produced by analysing a variety of data points according to a given formula. One metric used by some online services today is “Violative View Rate” (VVR), which is the percentage of all content views that were of content that is prohibited by community guidelines (some of which may be illegal).

response to complaints from UK users, for services to ensure their content moderation functions are designed to either:

- a) make an illegal content judgment in relation to suspected illegal content and, if it determines that content is illegal content, take the content down swiftly; or
- b) where a service is satisfied that its terms of service prohibit the types of illegal content defined in the Act which it has reason to suspect exist, consider whether the content is in breach of those terms of service and, if it is, take the content down swiftly.

19.16 We consider that complaints data derived from the second category would be an appropriate proxy for illegal content risk for the purposes of the proposed safety metrics.

19.17 Similarly, we consider that 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 service's terms of service do not vary significantly across the jurisdictions in which they operate. This means that, where a service is satisfied that the categories of content prohibited by its terms sufficiently cover priority illegal content when handling UK user complaints (in line with paragraph 19.15(b) above), non-UK user complaints resolved under those same categories during a test could also be used for the purposes of generating the proposed safety metrics.

19.18 We consider that this approach is preferable to prescribing sampling requirements within the measure to ensure a certain level of UK user representation in the on-platform test, as this could add to the costs of current practice and potentially limit the amount of data available to produce the safety metrics.

19.19 We recognise that using illegal content proxies from UK and non-UK complaints, rather than relying solely on data from illegal content judgments of UK user complaints, could lead to safety metrics being derived from content that is in breach of terms but doesn't necessarily amount to a priority offence. We nonetheless consider that this data qualifies as a useful "illegal content proxy" as it would give services an indication of how a design change might contribute to the distribution and possible virality of illegal content, which in turn contributes to the risk of users encountering that content through the deployment of a recommender system.

19.20 We do not envisage that the measure should apply in the context of design changes that:

- a) amount to a "significant change" and trigger the risk assessment duty under section 9(4) of the Act, as we intend to cover changes that are made more frequently and in the context of an ongoing practice of making smaller changes;
- b) are made in connection with a live and time-sensitive response to a national security risk or other emergency; or
- c) are not deployed for UK users of the service.

## Effectiveness

19.21 In considering the effectiveness of this measure, our starting point was the proposition, outlined in paragraph 19.3, that the way recommender systems are designed can influence the dissemination of illegal content on a service.



- 19.22 Based on the evidence Ofcom has gathered on the effectiveness of user reporting and on-platform testing<sup>584</sup>, we believe that services that follow this measure would gain better insights into the implications of design changes to their recommender systems on the dissemination of illegal content. These insights would enable services to avoid design choices which increase the likelihood of users encountering illegal content, thereby reducing the risk of harm to users.

#### Box D: Why on-platform tests?

While there are many ways of evaluating the impact of a recommender system, including through user surveys, sock puppet accounts and “debugging” exercises, we have chosen to focus this measure on ‘on-platform tests’ (e.g., A/B/x tests). Unlike some other evaluation techniques, these allow for direct causal inferences to be drawn (i.e., in this context, to see if a particular design choice could increase or decrease the dissemination of illegal content). On-platform tests take the form of randomised controls trials (RCTs) 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.

### Efficacy of the safety metrics

- 19.23 Our rationale for focusing on the safety metrics set out in Table 19.1 is as follows:

- a) **The total number of items identified as illegal content or as an illegal content proxy** – This metric would reveal to the service how many user complaints were upheld as illegal content or an illegal content proxy, and therefore how many unique items of this content were displayed to the control and treatment groups during an on-platform test. This would indicate the overall scale of risk to users in terms of the number of unique items of illegal content or appropriate proxies contained in the source pool that were surfaced by each variant of the recommender system.
- b) **The total number of impressions and reach per item of content identified as an illegal content proxy** – Services can use the first metric to produce a further metric that indicates the level of user exposure to illegal content across control and treatment groups. Impression data is important as it reveals how frequently a service’s user base may encounter illegal content or appropriate proxies. Reach, meanwhile, is important because it shows how many unique user accounts encountered illegal content or appropriate proxies. We consider that these metrics are relevant to an assessment of risk of users encountering illegal content, as they show the distribution of illegal content across users; for example, 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 witnessed several times by those users (high reach, low impression).

---

<sup>584</sup> Ofcom, 2023. Ofcom, 2023. [Evaluating recommender systems in relation to illegal and harmful content](#). [accessed 26 September 2023].

- 19.24 Once all metrics have been collected, the service would then be able to run a comparative analysis across all variants of the recommender system tested to evaluate the respective illegal content risk.
- 19.25 This measure would also include the recommendation that services keep a log noting the description of the design characteristics of each variant of the recommender system tested (for example, the respective features<sup>585</sup> and parameters<sup>586</sup>), and the safety metrics derived for each variant and a record of the design decision that was taken following the test. Maintaining a results log would help services understand how different variants of the recommender system (and its design characteristics) affects user safety, which in turn would allow service staff to make a more informed decision about which variants to fine-tune and deploy. As an outcome, services would be better equipped to diagnose and respond to design-based issues during the testing window, and fine-tune future alterations to their recommender system in the interest of user safety.
- 19.26 In addition, the results of the log could be used as an enhanced input in the form of data for future risk assessments a service might undertake as outlined in our draft Risk Assessment Guidance in Annex 5.

### Evidence relating to on-platform tests and logs of test results

- 19.27 We have also reviewed evidence relating to the effectiveness of conducting on-platform tests and keeping logs of those test results.
- 19.28 First, we understand that on-platform tests are an effective means of assessing how recommender system design choices can impact user safety. Research commissioned by Ofcom indicates that on-platform tests are one of the most robust evaluation methods for most, if not all, machine learning models<sup>587</sup> and are employed by many of the largest U2U services to understand the impact of their recommender system design choices for commercial outcomes such as clicks and views.<sup>588</sup>
- 19.29 Some services have openly disclosed using on-platform tests to examine how recommender systems could impact users:
- a) In 2020, LinkedIn explained in its engineering blog how it established inequality metrics to monitor barriers to economic opportunity for its users (e.g., in the form of exposure to job notifications). These inequality metrics were then able to be monitored during on-platform tests of product changes, including for LinkedIn's recommender system.<sup>589</sup>

---

<sup>585</sup> Features help the recommender system recognise content attributes (e.g., genre, subject, labels) and user characteristics (e.g., preferences) by giving context to data (i.e., features help the recommender make sense of data).

<sup>586</sup> Parameters are the internal settings that guide the recommender system translate features into practical content recommendations. They help the system understand the interrelationships/connections between content and users, enabling it to make effective recommendations.

<sup>587</sup> Recommender systems are underpinned by machine learning models.

<sup>588</sup> Ofcom, 2023. Ofcom, 2023. [Evaluating recommender systems in relation to illegal and harmful content](#). [accessed 26 September 2023].

<sup>589</sup> LinkedIn Engineering (Saint-Jacques, G., Sepehri, A., Li, N. and Perisic, I.) 2020. [Building inclusive products though A/B testing](#). [accessed 26 September 2023].

- b) In 2021, Twitter (now known as X) released the details of a large A/B/x test which considered how a proposed change to its recommender system altered the dissemination of political content in user feeds.<sup>590</sup>
- 19.30 We therefore consider that the use of safety metrics within on-platform tests is both technically feasible and an effective means of understanding and managing the illegal content risks associated with recommender system design.
- 19.31 Second, there is some evidence that services refer to a log of past test results when deciding what future changes to make to their recommender systems. Evidence obtained through engagement with Rumman Chowdhury<sup>591</sup>, an expert on recommender systems with extensive experience in algorithmic governance and on-platform tests, suggests that it is normal practice for services to maintain logs of test results, which detail the performance of recommender systems according to commercial metrics. Based on her experience, we noted that relevant teams typically have access to these results and that they are used to inform product changes as appropriate. Therefore, it is reasonable to conclude that the record of the safety metrics which we are proposing would be referred to by those same teams, making a log of those results an effective resource in the ongoing management of risk associated with recommender systems.
- 19.32 The measure requires services to consider the safety metrics obtained from on-platform tests when making design decisions to minimise risk to users. However, as the measure does not provide any obligation to implement prescriptive changes, the effectiveness of the measure would depend on the extent to which services act on the testing results they obtain.
- 19.33 We are conscious 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. However, this measure would not increase these risks, as it would not specify a recommendation for any new on-platform tests to be performed - only that services collect additional metrics within existing on-platform tests. We therefore do not believe that these considerations render the measure unethical or ineffective.
- 19.34 We acknowledge that smaller U2U services, that have fewer users and less content, may receive fewer relevant complaints that could be used to produce the necessary metrics. However, we consider that even a small number of complaints can uncover illegal content (and illegal content proxies), and indicate the extent of their dissemination (i.e., impressions and reach) during testing. In effect, there may be instances where a small number of complaints could be material to identifying a potentially risky design choice. In conclusion, even where complaints might be few and far between, they could contribute to services assessing the risk of a particular design choice. As results are recorded in a log over time, test results based on a small number of complaints may still contribute to cumulative awareness of risk across successive tests.

---

<sup>590</sup> Global Partnership on Artificial Intelligence (GPAI), 2022. [Transparency Mechanisms for Social Media Recommender Algorithms](#). [accessed 26 September 2023].

<sup>591</sup> Meeting with Rumman Chowdhury on Monday 20<sup>th</sup> February 2023.

## Costs and risks

- 19.35 There are significant costs associated with assembling and investing in new testing infrastructure. Doing so would require sophisticated computing infrastructure, specialist engineers, and maintenance costs – all of which would constitute a large capital investment.
- 19.36 If this measure was limited to services that already run on-platform tests when making changes to their recommender system, the cost of implementing this measure can be limited considerably. This is because such services would already have an established testing environment in place, as well as the specialist staff needed to execute on-platforms tests and implement the recommendations put forward in this measure. Additional costs in this scenario would be limited to:
- a) **Designing and setting up the new safety metrics (a one-off cost):** Setting up new safety metrics (as outlined in Table 19.1 above) would require services to identify the relevant data points, establish a data cleaning and preparation process, and establish a formula for analysing that data to produce the recommended metrics. This would require time from in-house data engineers and data scientists.
  - b) **Data storage (an ongoing cost):** Services would be required to collect and store additional data for the duration of the tests they perform. They would need to hold data relating to all the pieces of content that have been exposed to users in the treatment and control groups, including information about the classification of that content (deemed illegal or otherwise), and the number of impressions and reach of that content. Services would also need to maintain the log of past test results.
  - c) **Extended product management cycle (an ongoing cost):** It may take services additional staff time to review the new metrics that are produced as part of this measure, as well as to decide on how to act on them. This may require additional time commitment from in-house data engineers and product teams.
- 19.37 Regarding the first of these costs, Rumman Chowdhury explained to Ofcom that a new safety metric she had helped to establish required 2,000 human hours, or 55 weeks' worth of time, split between two in-house engineers and one researcher. This would be approximately £60,000 to £120,000.<sup>592</sup> We consider that the cost of establishing the safety metrics set out in this measure is likely to be towards the lower end of this range, or potentially below the range suggested. This is because we have specified how we would expect the metric to be constructed and we would expect services to already be capturing much of the required data (e.g., user complaints data). We also believe a one-off cost in the region of £60,000 or below is likely to be affordable for services who already employ on-platform testing.<sup>593</sup> This is because we expect services who have already implemented on-platform testing regimes are unlikely to include very small platforms for which this type of cost may be more difficult to cover.
- 19.38 Although, not directly relevant to this measure, we know that at least some of the larger services have experience in developing other types of safety metrics, which illustrates that some services have more general experience in this area. For example, YouTube produces a

---

<sup>592</sup> This is based on the wage assumptions we set out in Annex 14.

<sup>593</sup> In addition to this one-off cost, we assume an annual ongoing cost of approximately 25% of the one-off costs (i.e., £15,000), consistent with the assumption we have made elsewhere as described in Annex 14.

quarterly metric known as the ‘violative view rate’,<sup>594</sup> which involves reviewing a random sample of videos and assessing which breach YouTube’s Community Guidelines. Similarly, Meta produces a metric known as ‘prevalence’ to estimate the percentage of total views that were of content that breached Meta’s Community Standards across Facebook and Instagram.<sup>595</sup>

- 19.39 Regarding data storage costs, Rumman Chowdhury told Ofcom that considering the limited additional data that would require storage, the cost of the metric data would be negligible. This is especially true for the largest U2U services that already operate large data storage centers. Moreover, this data would not need to be retained beyond the duration of tests, which we understand do not typically run for more than several weeks, thus limiting data storage costs. The additional expense of storing the results log would be minimal, since this contains only aggregate, high-level information.
- 19.40 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 resource 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.
- 19.41 The impact of these costs may be lessened to some extent as identifying and addressing potential causes of harm upfront may reduce the costs services incur mitigating harm after the fact. For example, reducing the extent to which recommender algorithms disseminate illegal content may reduce the costs content moderation teams incur dealing with reports of illegal content.
- 19.42 Altogether, our analysis indicates that the largest U2U services that already perform on-platform tests would be able to meet the costs of this measure. We also consider the same is likely to 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. We believe the other two costs would also be affordable for smaller services as the costs are likely to strongly correlate with their size and the amount of on-platform tests they already do. For example, a smaller service is likely to require less data storage (given fewer users and content volume on their sites) and have fewer test results to analyse (given we would expect them to run fewer tests).

## Rights impacts

### Freedom of expression

- 19.43 This mitigation focusses on generating organisational risk awareness from which safety-conscious design decisions may be made; it does not recommend that a service make a particular design decision based on results of the new testing safety metrics. To the extent that there was any indirect impact on the right to freedom of expression, we would consider

---

<sup>594</sup> YouTube (O’Connor, J.), 2021. [Building greater transparency and accountability with the Violative View Rate](#). [accessed 26 September 2023].

<sup>595</sup> Meta, 2022. [Prevalence](#). [accessed 26 September 2023].

it proportionate in pursuit of a legitimate aim, since it would be to reduce users' exposure to illegal content.

## Privacy implications

19.44 We believe that the implementation of this measure may have an impact on the right to privacy in two ways:

- a) The production of additional safety metrics would require services to collect, store, and analyse complaints and content. For example, the complaints handling process could involve the processing and storage of personal data of individuals who posted the content, who are identifiable in the content, or who submitted a complaint.
- b) On-platform testing involves the random allocation of users into control and treatment groups, which raises a question of user consent to the processing of new personal data or processing existing data for new purposes. While publicly available information indicates that Facebook, Instagram, Twitter, and LinkedIn carry out frequent on-platform tests, less is known about how user consent is managed. Meta (Facebook<sup>596</sup> and Instagram<sup>597</sup>) set out in their terms of service that tests are carried out as part of research and innovation, though do not specify what type of tests are run and how often. Twitter has more detailed information on its testing programme<sup>598</sup> and has published a blog on how research and experimentation is conducted. Due to the frequency of on-platform tests and the high number of participants, we believe that users are not specifically notified that they are part of an on-platform test and for the most part, consent would be obtained alongside consent to the terms of service more broadly.

19.45 We consider any implications for the right to privacy is justified by the importance of on-platform testing as a robust method of identifying and managing the risk of recommender systems exposing users to illegal content. While other testing types may not involve the processing of personal data (e.g., user surveys whose results are anonymised), they are less effective for the purposes of evaluating illegal content risk.

## Who the measure would apply to

19.46 We have considered to which services it would be appropriate to apply the measure outlined above.

19.47 As part of our proportionality assessment, we have considered the following options for application to U2U services:

- Option A: All U2U services to develop and carry out on-platform testing of their recommender systems; or
- Option B: Only those U2U services that already employ on-platform testing of their recommender systems should extend these tests to observe specific safety metrics; or
- Option C: Only those U2U services that both (i) already employ on-platform testing of their recommender systems and (ii) have assessed that they are high or medium

---

<sup>596</sup> Facebook (Meta), 2023. [Terms of Service](#). [accessed 26 September 2023].

<sup>597</sup> Instagram (Meta), 2023. [Data Policy](#). [accessed 26 September 2023].

<sup>598</sup> X (formally Twitter), 2017. [About the X Experiments Programme](#). [accessed 26 September 2023].

risk for at least two types of the illegal harms identified in paragraph 19.53 in their latest illegal content risk assessment.

19.48 For the reasons summarised below, our provisional view is option c is the most proportionate.

### Type of provider (do they already undertake on-platform testing)

19.49 As explained above, the costs of establishing testing infrastructure from scratch are significant. This would involve building a virtual testing environment and onboarding or training specialist staff to run these tests.

19.50 We therefore provisionally think it would only be proportionate for this measure to apply to services that already run some form of on-platform testing. Box D above explains why we are focusing on on-platform testing versus other methods that could be used to evaluate recommender systems.

### Type of provider (associated risk)

19.51 For those services that already run on-platform tests on their recommender systems, and have testing infrastructure in place, establishing new safety metrics as part of those tests would still entail some costs irrespective of the size of the service.

19.52 We therefore provisionally think this measure should not apply to services whose risk assessment indicates that users face a low risk of encountering harms relevant to recommender systems. In such cases, the required outlay of the measure would be disproportionate considering the risks it is intended to mitigate.

19.53 As such, we consider that this proposed measure should only apply to services that have identified a high or medium risk for at least two of the following priority offences, which are the kinds of offences that we consider recommender systems could affect (please see the relevant Ofcom's Register of Risks chapters in Chapter 6):

- 6B: Terrorism offences;
- 6C: Child Sexual Abuse Material (CSAM);
- 6D: Encouraging or assisting suicide (or attempted suicide) or serious self-harm
- 6E: Harassment, stalking, threats and abuse;
- 6F: Hate offences;
- 6H: Drugs and psychoactive substances.
- 6L: Extreme pornography offences;
- 6M: Intimate image abuse offences);
- 6P: Foreign interference offence;

### Provisional conclusion

19.54 The way content recommender systems are designed can influence the extent to which certain categories of illegal content are disseminated on a service. As such, design changes to recommender systems can influence how much illegal content is served to end users. Understanding better how changes to recommender systems can help services reduce the



extent of this. As such, we provisionally conclude that this proposed measure would have significant benefits for user safety. By gathering and then consulting information about the impact of recommender system design changes on the dissemination of illegal content, services would be in a better position to make safety-conscious design choices than they otherwise would. All else being equal, this should improve outcomes for users by reducing the risk of harm from illegal content they may encounter on a service.

- 19.55 The measure would result in additional costs for services, including those borne from establishing the new safety metrics, requiring more data storage, and extending product management cycles. Given that we are focusing this measure on services that already conduct on-platform tests, our provisional view is that for services in scope of the measure these costs are likely to be a relatively small addition to their existing on platform testing costs. These services would not be asked to run any new tests, nor to establish any new testing infrastructure. In view of this and considering the measure confer would important benefits, we consider that our proposal would be proportionate.
- 19.56 We also propose to recommend the measure only for services that have identified a high or medium risk for at least two of the illegal harms identified in paragraph 19.53 above in their latest illegal content risk assessment. The benefits from the measures are likely to be greater for such services.
- 19.57 We propose to recommend in our Codes of Practice on Terrorism, CSEA and other duties that relevant U2U services should observe safety metrics whenever they carry out on-platform tests on their recommender systems to understand whether a proposed design change would increase the dissemination of illegal content.
- 19.58 As part of this measure:
- a) Services should produce the safety metrics outlined in Table 19.1 (or equivalent), including the total number of items identified as illegal content or an illegal content proxy in response to user complaints during testing, and the total number of impressions and reach per item. This data should be derived from a service's complaint handling procedures.
  - b) Services should keep a log of the test results, noting the safety metrics produced in respect of each variant of the recommender system tested. For each variant of the recommender system being tested, there should be a clear description of the respective design characteristics being evaluated. Following the test results, it should be noted which variant was deployed (i.e., the design decision taken forward).
  - c) Staff involved directly or indirectly in the development and testing of recommender systems and should consult the log in the context of future design changes to their recommender system.

## 20. Enhanced User Control (U2U)

### What is this chapter about?

In this chapter we explore features that U2U services can use to help users manage the risk of being exposed to illegal content. These measures are aimed at giving users more control or understanding of the content they encounter and allowing them to make judgements about the risk of encountering illegal content.

### What are we proposing?

We are making the following proposal for all large services that identify as medium or high risk for any of the specified harms listed at the following footnote,<sup>599</sup> have user profiles and have at least one of the functionalities listed at the following footnote:<sup>600</sup>

- **Services should offer every registered user options 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.**

We are making the following proposal for all large services that identify as medium or high risk for any of the specific harms listed at the following footnote<sup>601</sup> and enable users to comment on content:

- **Services should offer every registered user the option of disabling comments on their own posts.**

We are making the following proposal for all large services that identify as medium or high risk of fraud or foreign interference, and already operate a notable user verification scheme and/or monetised user verification scheme:

- **Services should have, and consistently apply, internal policies for operating these schemes and improve public transparency for users about what verified status means in practice.**

### Why are we proposing this?

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 our proposals are sufficient to justify the costs we have identified. There is a

---

<sup>599</sup> Coercive and controlling behaviour; harassment, stalking, threats and abuse; hate; grooming; encouraging or assisting suicide or serious self-harm.

<sup>600</sup> User connections; posting content; or user communication (including but not limited to direct messaging and commenting on content).

<sup>601</sup> Harassment, stalking, threats and abuse; hate; grooming; or encouraging or assisting suicide or serious self-harm.

degree of uncertainty about some of the costs. In order to ensure that we are acting proportionately, we are proposing to target the measures at medium or high-risk large services.

Our evidence suggests that some users pay attention to verified status of accounts when deciding whether to engage with and trust content. If users do not understand what verified status conveys, there is a risk that they could succumb to impersonation fraud or disinformation disseminated by a hostile foreign state actor. Our proposed measure regarding verification schemes addresses this risk.

### What input do we want from stakeholders?

- Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.
- Do you think the first two proposed measures should include requirements for how these controls are made known to users?
- Do you think there are situations where the labelling of accounts through voluntary verification schemes has particular value or risks?

## Introduction

---

- 20.1 In this chapter we explore features and functionalities that can be used to assist users in managing the risk of illegal harm. These measures are aimed at giving users more control or understanding of the content they encounter, allowing them to make judgments about the risk of encountering illegal content, and operate on the service based on their judgement and experience.
- 20.2 Our recommendations include functionalities that allow users to manage risk by, for example, becoming uncontactable, disabling comments, and blocking other users. We are also recommending that services apply good design practices when offering features like user verification schemes, so that they support users in identifying content from verifiable notable figures and distinguishing between real and fake accounts.
- 20.3 The tools recommended in this phase of work on illegal content are distinct from the user empowerment duties for category 1 services; those will require designated services to provide features specifically for adult users around *legal* content relating to suicide, self-injury and eating disorders, and optional identity verification schemes. We will consult on these requirements in later phases of our work.

## Measure to give all users the ability to block and mute other user accounts

---

### Harms that the measure seeks to address

- 20.4 The ability of users to engage with one another on U2U services, and the sheer extent of these potential interactions, carries risk that users may encounter priority illegal content shared by others. Indeed, the Alan Turing Institute has estimated that 10-20% of people in the UK have been personally targeted by abusive content online.<sup>602</sup> Similarly, in research conducted by Ofcom on Video Sharing Platforms (VSPs), one third of users said they have

---

<sup>602</sup> The Alan Turing Institute (Margetts, Z and Harris, A), 2019. [How much online abuse is there?](#) [accessed 8 September 2023].

witnessed or experienced hateful content.<sup>603</sup> In that study, 59% claimed to have been exposed to contact harms in the last three months, with unwelcome friend requests or messages and trolling being the most commonly experienced (40% and 36% respectively).<sup>604</sup> Research by the Victims Commissioner found that, among people who had experienced online abuse, on average they had experienced multiple types of abuse per person. 63% of this abuse came from strangers.<sup>605</sup>

- 20.5 We recognise that people reporting seeing abusive or hateful content does not necessarily mean it is illegal, though it is likely that a user would encounter illegal content through similar channels. Users may encounter this content through a variety of functionalities and channels on these services, as part of user communication.<sup>606</sup> Illegal content could be shared directly, such as through a messaging functionalities or a comment on content. Or it could be shared publicly, such as through a post shared and distributed on a newsfeed. The perpetrator could be either known or unknown to the user.
- 20.6 These communications functionalities present a particular risk of certain illegal harms, as drawn out in the evidence and rationale below. These include controlling or coercive Behaviour; hate; harassment, stalking, threats and abuse; encouraging or assisting suicide or serious self-harm; and grooming.
- 20.7 As outlined in Volume 2: Chapter 6 - Part 1, **direct messages** are often used as a means of sharing illegal content or targeting victims with other illegal activity:
- a) As outlined in the Volume 2: Chapter 6E Harassment, stalking threats and abuse offences, paragraph 6E.32, evidence suggests that children are more vulnerable to these harms. Children reported that receiving unsolicited messages from strangers is very common online and particularly among older girls, with evidence that children were receiving explicit content and racist abuse.<sup>607</sup> One study found that 76% of girls aged 12-18 had been sent unsolicited nude images of boys or men, including repeatedly in some cases.<sup>608</sup>
  - b) The ability to communicate directly with children online may be exploited to commit or facilitate grooming offences. As identified in the Volume 2: Chapter 6C CSEA (grooming and CSAM), paragraph 6C.90, direct messaging is frequently used as a means of initiating contact and developing relationships with victims online through frequent communication in a private setting. The ability to share images in this setting can also enable child sexual abuse offences, such as persuading children to self generated indecent images (SGII) Volume 2: Chapter 6C CSEA (grooming and CSAM), paragraph 6C.85 and 6C.86.
  - c) The Suzy Lamplugh Trust found that, of those who experienced stalking, the proportion of those who experienced stalking on social media services jumped from 59% before the

---

<sup>603</sup> Ofcom, 2021. [User experience of potential online harms within video-sharing platforms: a report from Yonder](#). [accessed 8 September 2023].

<sup>604</sup> Contact harms are potential online harms that originate from the behaviour of other users.

<sup>605</sup> Victims Commissioner, 2022. [The Impact of Online Abuse: Hearing the Victims' Voice](#). [accessed 25/08/2023].

<sup>606</sup> We define User Communication as “functionalities by means of which users can communicate with one another either synchronously or asynchronously. Includes communication across open and closed channels.” From Ofcom Risk Assessment Glossary.

<sup>607</sup> Ofcom, 2022. [Children's Media Lives](#). [accessed 8 September 2023].

<sup>608</sup> UCL, 2021. [Tackling image-based sexual harassment and abuse](#). [accessed 8 September 2023].

Covid-19 pandemic to 82% afterwards.<sup>609</sup> Texts or direct messages were the most common element of digital stalking behaviour, experienced by 65% of participants. Threats were also commonly made via online digital communication. This also increased during the Covid-19 pandemic, experienced by 27% of participants before the first lockdown, and 47% after it.<sup>610</sup>

- d) Across all adults, Ofcom's Online Experiences Tracker found that 4% of respondents had received unwanted/unsolicited sexual or nude images in the past four weeks.<sup>611</sup> This is discussed in more detail in Volume 2: Chapter 6S 'Cyberflashing' offence.

20.8 As outlined Volume 2: Chapter 6E(Harassment, stalking threats and abuse offences), paragraphs 6E.69 and 6E.70, functionalities allowing users to comment on content may provide another avenue through which illegal content can be shared and encountered:

- a) A 2014 study by Pew Research found that 22% of internet users had been a victim of online harassment in the comments section of their uploads.<sup>612</sup>
- b) 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.<sup>613</sup>
- c) 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.<sup>614</sup>
- d) Grooming and CSAM offences may be facilitated through the use of comments, which have been used by perpetrators not just to communicate directly with children, but also as a means of engaging in sexualised conversations and, in the context of livestreaming, inciting them to engage in sexual activity in real time, which can then be screen-recorded without consent (see Volume 2: Chapter 6C CSEA (grooming and CSAM), paragraph 6C.85-6C.87).
- e) Perpetrators of online cyberflashing can use channels such as direct messaging and commenting to harass victims by sharing unsolicited explicit images as outlined in paragraph 21.7d.
- f) As outlined in the Volume 2: Chapter 6D Encouraging or assisting suicide (or attempted suicide) or serious self-harm offences, paragraphs 6D.67-6D.69 functionalities allowing users to **comment on content** may also be used to share content which encourages suicide, or to encourage the suicide depicted in a piece of content such as livestream.

20.9 Beyond the risk of encountering illegal content itself, certain service functionalities may allow users to interact in a way that facilitates the commission of a priority offence, such as through their use of user profiles. Our Volume 2: Chapter 6 - Part 1 discusses the risk posed by users being able to create multiple and anonymous or fake accounts in the context of

---

<sup>609</sup> Suzy Lamplugh Trust, 2021. [Unmasking stalking a changing landscape](#). [accessed 7 September 2023].

<sup>610</sup> Suzy Lamplugh Trust, 2021.

<sup>611</sup> Ofcom, 2020. [Internet users' experience of potential online harms: summary of survey research](#). [accessed 8 September 2023].

<sup>612</sup> Pew Research Centre (Duggan, M), 2014. [Online Harassment](#). [accessed 29 August 2023].

<sup>613</sup> YouTube, 2022. [YouTube Community Guidelines enforcement – Google Transparency Report](#). [Accessed 25 August 2023].

<sup>614</sup> Ofcom, 2022. [Online Experiences Tracker Data tables waves 1 and 2](#). [accessed 19 July 2023].

harassment, stalking, threats and abuse, encouraging or assisting suicide (or attempted suicide) and serious self-harm, and grooming. For example:

- a) Individuals who stalk, harass or threaten online are known to sometimes run multiple accounts when interacting with their victims, so that in expectation of at least one account being reported and banned, they can seamlessly move to another.<sup>615</sup>
- b) 26% of women surveyed by Refuge reported being messaged by an account they suspected to be fake.<sup>616</sup>
- c) We are aware of cases where perpetrators have created multiple profiles to maliciously encourage others to take their own lives.<sup>617</sup> Further information can be found in the Volume 2: Chapter 6D Encouraging or assisting suicide (or attempted suicide) or serious self-harm offences.
- d) Grooming perpetrators have been known to create ‘fake’ user profiles to present in a less threatening way to children, including in the case of a perpetrator posing as a teenage girl online to groom and sexually abuse 500 boys online.<sup>618</sup> In addition, user networking features such as connections can be used by perpetrators in a “scatter-gun” approach to contact a large number of children, and the visibility of user connections can be used to create trust where a perpetrator has a number of mutual connections with a child Volume 2: Chapter 6C CSEA grooming and CSAM), paragraphs 6C.75-6C.77.

20.10 We understand that some users can be at greater risk of being targeted with illegal content from other users based on certain characteristics of the user, for example, because they have a certain protected characteristic, status or profession:

- a) Ofcom research found that women are particularly likely to be negatively affected by hateful, offensive or discriminatory content and trolling online.<sup>619</sup> Similarly, Reddit found that accounts perceived as women received a higher rate of hateful content in direct messages.<sup>620</sup> And among women surveyed by Refuge, 25% had abusive or upsetting content shared with them.<sup>621</sup>
- b) Diaspora communities may be particularly targeted by harmful and potentially illegal content from other accounts, in circumstances that could raise harassment concerns. Recently, it has been reported that certain groups reporting on China have faced online trolling and Chinese-state-led intimidation in the UK, including Hong Kong and Uighur diaspora groups.<sup>622</sup>
- c) High profile users, such as celebrities or other public figures, may be at an elevated risk of being targeted with illegal harms. England football players Marcus Rashford, Jadon

---

<sup>615</sup> UK Home Office, 2021. [Anonymous or multiple account creation: improve the safety of your online platform](#). [accessed 29 August 2023].

<sup>616</sup> Refuge, 2021. [Unsocial Spaces](#). [accessed 30 August 2023].

<sup>617</sup> Phillips, J G., Diesfeld, K. and Mann, L., 2019. [Instances of online suicide, the law and potential solutions](#), (p.8), *Psychiatry, Psychology and Law*, 26 (3). [accessed 27 January 2023].

<sup>618</sup> BBC News, 2021. [David Wilson: Sex offender who posed as girls online jailed for 25 years](#), *BBC*, 10 February [accessed 11 August 2023].

<sup>619</sup> Ofcom, 2022. [Online Nation](#) [accessed 8 September 2023].

<sup>620</sup> Reddit, 2022. [Reddit’s Prevalence of Hate Directed at women](#). [accessed 8 September 2023].

<sup>621</sup> Refuge, 2021. [Unsocial Spaces](#). [accessed 30 August 2023].

<sup>622</sup> Freedom House (Datt, A., and Dunning, S.), 2022. [Beijing’s Global Media Influence 2022](#), Country Reports: United Kingdom [accessed 17 February 2023]; and Hope, C., 2021: [Exclusive: Uighers harassed and abused by Beijing in UK, minister admits](#), *The Telegraph*, 13 March. [accessed 17 February 2023].

Sancho and Bukayo Saka were targeted with racist abuse online in the aftermath of the Euro 2020 final. This abuse occurred through posts, comments and tweets on both Instagram and X (formerly known as Twitter).<sup>623</sup> More recently, Brentford footballer Ivan Toney was sent racist abuse via a direct message on Instagram.<sup>624</sup> Following these incidents, the Alan Turing Institute and Ofcom produced a broad report into footballer abuse that found that around one in twelve personal attacks targeted a victim's protected characteristic, such as their race or gender.<sup>625</sup> The same report into Footballer Abuse found that hundreds of abusive tweets are sent to Premier League footballers every day.<sup>626</sup>

- d) A three-year global study on gender-based online violence against women journalists, reported by UNESCO, found that nearly three-quarters of a sample said they had experienced online violence in the course of their work. Threats of physical violence, including death threats, were identified by 25% of the respondents, and threats of sexual violence were identified by 18%.<sup>627</sup>

20.11 While some of this content may not amount to illegal content, the scale of 'negative' content demonstrates the risk of potentially illegal content that certain users are at risk of encountering.

## Options

20.12 The functionalities identified above that present a risk of encountering illegal content are valuable to many users and are often integral to the operation and business model of most U2U services. As such, we do not consider it proportionate to recommend that services remove these functionalities in order to protect people from illegal harms.

20.13 However, the evidence set out in Volume 2: Chapter 6 - Part 1, user base risk factors shows how users are often in a position to know that they will be or have been targeted, and in some cases to predict who they will be targeted by. This suggests that tools that enable users to block or mute other users, whom they consider to present a risk of harm, could provide an effective means of mitigating the harms outlined above from arising between users. It follows that, in order for one user to be able to block or mute other users, those users would need to be identifiable, for example by having a user profile. With that in mind, we have considered the following options.

20.14 In these options, based on industry practice we consider the following terms to have the following meanings, where "user A" wishes to protect themselves against "user B" (where "user B" is an identified user or a non-connected user):

- a) "**Block**", which is commonly-used terminology to refer to a user tool widely provided by U2U services, which can be used to block individual users or non-connected users globally. Here, it means that:

---

<sup>623</sup> Landler, M., 2021. [After Defeat, England's Black Soccer Players Face a Racist Outburst](#), *New York Times*, 12 July. [accessed 17 February 2023].

<sup>624</sup> Sky Sports, 2022. [Ivan Toney: Brentford striker contacted by police after racist Instagram message](#), *Sky*, 15 October. [accessed 8 September 2023].

<sup>625</sup> Ofcom, 2022. [Crossing the line: Seven in ten Premier League footballers face Twitter abuse](#). [accessed 7 September 2023].

<sup>626</sup> Ofcom, 2022.

<sup>627</sup> International Centre for Journalists, 2022. [The Chilling: A global study of online violence against women journalists](#). [accessed 8 September 2023].



- i) User B cannot send direct messages to user A and vice versa;
  - ii) 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 by user B; and (2) content originally posted by user B which is subsequently posted by another user; and
  - iii) User A and user B, if they were connected, will no longer be connected.
- b) “Mute” is similarly commonly-used terminology to refer to a user tool widely provided by U2U services, and here means that user A will not encounter any content posted by user B on the service, including: (1) reactions to and ratings of content by user B; and (2) 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. Muting is softer than blocking, for example if user B were only muted then they would still be able to direct message user A.

20.15 The options we consider here are:

- a) Option 1 (individual account blocking and muting): U2U services should provide all users with the ability to block and mute other user accounts individually. We expect that this would be a feature in the settings of a user account. This is already an industry standard safety option offered by platforms to users as a means of keeping themselves safe.
- b) Option 2 (global blocking of any non-connected account): U2U services should provide users with a public user profile a clear and accessible means of making themselves uncontactable to users without mutually validated connection. The fact that the user is on the service could still be seen by other users, but the content of their profile would not be visible except to connected accounts.<sup>628</sup> For the user it would functionally operate as though all non-mutually validated connected users were blocked.

## Effectiveness

20.16 We believe that these options would protect users from illegal content, because a user who could predict that they would be targeted with illegal content either by a specific person (e.g. a harasser) or by people in general (either because of a particular characteristic they had, or because of an harasser who created multiple accounts) would be able to prevent it from happening. Individual users could also calibrate their levels of protection based on their own assessment of the levels of harm to them that could occur.

20.17 Blocking would be powerful and would work for all users, including for those who may be more vulnerable to encountering illegal content or behaviours, have an overall lower risk appetite, or who have experienced potentially illegal content from a particular user and wish to protect themselves in future. Blocking would also operate as a safety net in the likelihood that even the most comprehensive moderation tools may be unable to identify and remove all illegal content.

20.18 Muting would be a softer tool than blocking but could also allow users to control their risk of encountering illegal content from other users in situations where they may not wish to go as far as to block that account. For instance, a user may wish to maintain an online connection with another account, such as that of a family member or ex-partner, but ‘mute’ them to reduce the risk of indirectly encountering illegal content from the account in the future

---

<sup>628</sup> “Seen” in this context is relevant to identification via either a User Profile or User identification.

whilst ensuring that the connection was not aware that this action had been taken. For example, Refuge found that 15% of the women survivors responding to the survey said the abuse worsened when they reported the perpetrator or took an action to mitigate the abuse, such as blocking the perpetrator online.<sup>629</sup>

## Discussion of option 1 (individual account blocking and muting across certain harms)

- 20.19 Young users have identified this measure as important across a variety of studies. In research conducted into online hate, unwanted sexual content, and appearance-related content, the University of Surrey found that participants endorsed developing and practising digital skills like “blocking the content and the users who create and share the content.”<sup>630</sup> While these content types may not amount to illegal content, it nonetheless suggests that young users consider user blocking to be a useful tool to reduce exposure to harmful and potentially illegal content.
- 20.20 In a study from Thorn in 2021 that looked at self-generated indecent imagery (SGII) and young users,<sup>631</sup> 66% of respondents responded to a harmful online experience by blocking the user, and 27% muted the user.<sup>632</sup> Blocking was much more common than reporting among the young users included in the report. The study also found that the frequency of blocking is higher among teenage girls; whereas 76% of the 13-17 year old girls in their sample had blocked a user following a potentially harmful online experience, only 56% of teenage males in the sample did so.<sup>633</sup> These figures are similar to Ofcom research showing that two-thirds of 12-17 year-olds indicated that they had blocked someone on social media that they did not wish to hear from.<sup>634</sup> This suggests that providing child users with the option to block or mute users would be an effective tool to enable them to protect themselves online, particularly in relation to grooming and other child sexual abuse offences where the child has experienced behaviours that make them feel unsafe or uncomfortable. We recognise that many children may find it difficult to end contact with users online, and therefore propose to recommend in Chapter 18 that services provide child users with supportive information at the point of blocking or muting. This may increase the efficacy and use of this measure in that context.
- 20.21 The Crime Survey for England and Wales found that 5.7% of respondents over the age of 16 had experienced stalking with an online element.<sup>635</sup> Victims of cyberstalking adopt various defence strategies. Research suggests that one of the most effective methods is to block communications from stalkers.<sup>636</sup>

---

<sup>629</sup> Refuge, 2021. [Unsocial Spaces](#) [accessed 4 September 2023].

<sup>630</sup> Young users in this context were defined as people aged 13 to 21 in England. Source: Setty, E., 2022. [Young People’s Perspectives on Online Hate, Unwanted Sexual Content, and ‘Unrealistic’ Body- and Appearance-Related Content: Implications for Resilience and Digital Citizenship](#). *Youth*, 2 (2). [accessed 2 March 2023].

<sup>631</sup> An indecent image of a person under 18 years is CSEA illegal content.

<sup>632</sup> Thorn, 2021. [Responding to Online Threats: perspectives on Disclosing, Reporting, and Blocking](#) [accessed 4 September 2023].

<sup>633</sup> Thorn, 2021. Note that the sample size is 1000.

<sup>634</sup> Ofcom, 2023. [Children’s Media Use and Attitudes](#). [accessed 22 August 2023].

<sup>635</sup> Office of National Statistics, 2022. [Stalking: findings from the Crime Survey for England and Wales - Office for National Statistics \(ons.gov.uk\) Year ending March 2022 edition](#) [accessed 8 September 2023].

<sup>636</sup> Tokunaga, R. S. and Aune, K. S., 2017. [Cyber-Defense: A Taxonomy of Tactics for Managing Cyberstalking](#). *Journal of Interpersonal Violence*, 32 (10). [accessed 8 September 2023].

- 20.22 Some civil society also supports blocking and muting features as a means of protecting users. Journalist groups such as Pen America state that, for vulnerable users like journalists and writers, “limiting contact with an abusive account and limiting exposure to abusive content—via features like blocking, muting, and restricting— can help... protect ... from unwarranted, inappropriate, or harmful conduct.”<sup>637</sup>
- 20.23 There is evidence that tools allowing users to block and mute other users is standard practice across some larger U2U services. Each of X, Facebook, Instagram, TikTok, LinkedIn, Snapchat, YouTube, Medium, WordPress, Tumblr and Reddit allow for user blocking and/or muting in some form.<sup>638</sup> Meta’s latest service, Threads, allows users to “unfollow, block, restrict or report a profile on Threads ... and any accounts ... blocked on Instagram will automatically be blocked on Threads.”<sup>639</sup> Payment providers such as PayPal<sup>640</sup> and digital marketplaces such as eBay<sup>641</sup> also provide user blocking tools.
- 20.24 In response to our 2022 Illegal Harms Call for Evidence, several stakeholders flagged the value and impact of existing user blocking and muting tools, or called for their implementation:
- a) The Anti-Defamation League noted that individual user blocking is one of the 32 ‘design patterns’ in their “[Social Pattern Library](#)”<sup>642</sup> that work to mitigate the presence of online hate and harassment, and should be enhanced and deployed more widely by industry.<sup>643</sup>
  - b) End Violence Against Women Coalition, Glitch, Refuge, Carnegie UK, NSPCC, 5Rights and Professors Clare McGlynn and Lorna Woods, represented by Glitch in their response, suggest that services provide users with tools to block or mute users, or categories of user. Glitch emphasised that more user tailoring tools are key to enhancing users’ experiences online in the context of offences such as cyberstalking, harassment, hate and coercive controlling behaviour<sup>644 645</sup>.
  - c) Chayn recommended “customisable settings that allow users to: control how their images and other media can be downloaded and shared, who can be in touch with them and how.”<sup>646</sup>
  - d) Global Partners Digital advocated for allowing users to block content from particular people or groups or on particular topics, or content from unverified or anonymous accounts, such as X’s (at the time, Twitter’s) Block Party tool.<sup>647</sup>
- 20.25 We recognise that the effectiveness of individual blocking tools may be limited in circumstances where the blocked user creates new accounts through which to continue

---

<sup>637</sup> Pen America, [Online harassment Field Manual; Blocking, Muting and Restricting](#). [accessed 12 September 2023].

<sup>638</sup> Pen America, [Online harassment Field Manual; Blocking, Muting and Restricting](#). [accessed 6 June 2023].

<sup>639</sup> Meta, 2023. [Introducing Threads: A New Way to Share With Text](#). [accessed 17 July 2023].

<sup>640</sup> Paypal. [How Do I Block Another PayPal User](#). [accessed 12 September 2023].

<sup>641</sup> eBay. [Blocking a buyer on eBay](#). [accessed 12 September 2023].

<sup>642</sup> ADL. [Social Patterns Library](#). [accessed 8 September 2023].

<sup>643</sup> ADL response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

<sup>644</sup> Glitch response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

<sup>645</sup> Carnegie United Kingdom Trust (The End Violence Against Women Coalition, Glitch, Refuge, Carnegie UK, NSPCC, 5Rights, McGlynn, C and Woods, L), 2022. [Violence Against Women and Girls \(VAWG\) Code of Practice](#). [accessed 12 September 2023].

<sup>646</sup> Chayn response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

<sup>647</sup> Global Partners Digital response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

targeting the blocking user, as described in paragraph 20.16 above. We nonetheless consider that the ability to block individual users is an effective tool to help users manage their risk online, and any residual risk associated with the practices of determined users may be addressed through global blocking tools, as considered below in relation to Option 2.

## Discussion of option 2 (global blocking of any non-connected account)

- 20.26 We consider that global blocking of non-connected accounts may address the risk of illegal content for users in various circumstances, including users that are high profile or public figures, or those who are being targeted by a blocked user via new accounts. Enabling users to make themselves inaccessible to non-connected accounts<sup>648</sup> has the advantage, beyond blocking at an individual level, of pre-empting illegal harm before it occurs.
- 20.27 As outlined in paragraph 20.10, certain types of user are particularly at risk of harm from non-connected accounts, such as celebrities with a large social media presence who are vulnerable to online abuse and stalking. This content might be directed to them via direct messages, but also comments on posts. Sometimes hateful content sent to an individual through a comment functionality can be amplified by the scale of comments that individual receives. Ofcom research on footballer abuse on X (then known as Twitter) suggests users may send just one abusive comment to an individual,<sup>649</sup> but sometimes that targeted individual can receive comments from many users simultaneously.
- 20.28 It also addresses the risk of harm to child users in circumstances where, as outlined in paragraph 20.7(b), they are targeted by non-connected accounts in an attempt to initiate communication with the intent to groom. We recognise that the application of a blocking feature would overlap to some extent with our proposed recommendation in Chapter 18<sup>650</sup> that, by default, non-connected accounts should not be able to send direct messages to child users. However, we consider that the ability to separately block users individually or using a global blocking feature would be effective in reducing the risk of grooming. This is because it would remove a user's ability to see all content posted by blocked and non-connected users. For example, a potential perpetrator could be posting content that gives the impression they are under 18, which in turn may increase the likelihood that a real under 18 user accepts a direct message request or a request to connect. Adding the option to block unconnected user content via this message would reduce the risk of grooming occurring in this scenario.
- 20.29 This option also addresses the residual risk from individual user blocking; that an offender once blocked sets up new accounts to continue contact with a victim. As outlined in paragraph 20.25, this pattern of malicious and repeated targeting of users through the creation of multiple or anonymous and fake accounts has been observed in technology-enabled domestic abuse, stalking and coercive control,<sup>651</sup> as well as encouraging or assisting suicide.<sup>652</sup> The ability to global block non-connected users may therefore assist in those circumstances.

---

<sup>648</sup> As described in the options section, this translates into an easy to use setting/option that makes them unable to be contacted across any functionality by users that are not a mutually validated connection.

<sup>649</sup> Ofcom, 2022. [Crossing the line: Seven in ten Premier League footballers face Twitter abuse](#). Pg 30 [accessed 7 September 2023].

<sup>650</sup> ADD REF TO CHAPTER.

<sup>651</sup> Refuge, 2021. [Unsocial Spaces](#). [accessed 30 August 2023].

<sup>652</sup> Phillips, J G., Diesfeld, K. and Mann, L., 2019. [Instances of online suicide, the law and potential solutions](#), *Psychiatry, Psychology and Law*, 26 (3). [accessed 27 January 2023].

- 20.30 We have evidence that some services currently provide users with a functionality allowing them to globally block all non-connected accounts:
- a) At the time of writing, X gives users the option to set replies only to those accounts that the user follows.<sup>653</sup>
  - b) By default, Snapchat users under 18 must opt-in to being friends in order to start chatting with each other.<sup>654</sup>
  - c) Instagram allows for a more technically complex pre-emptive blocking of new accounts of the user of a blocked account.<sup>655</sup> It also enables users to block all direct messages.
  - d) Facebook’s contact rules limit messaging to friends and other opt in associations, including Facebook dating and marketplace. Users outside of these categories cannot send a message; they can only send a request to message.<sup>656</sup>
  - e) Discord allows users to block direct messages from users on a server that are not on their friends list.<sup>657</sup>
- 20.31 As outlined in paragraph 20.24, stakeholders responding to Ofcom’s 2022 Illegal Harms Call for Evidence supported the use and further deployment of blocking tools, including global blocking.
- 20.32 The efficacy of any pre-emptive blocking would likely depend in part on the ease of use and prominence of the feature. Evidence from behavioural experiments on new system features for safety and privacy has shown that users may not use a new or novel feature due to possible lack of familiarity.<sup>658</sup> However, we reason that this feature would be so similar to user blocking and muting, an already widespread feature, that users are highly likely to understand it and more likely to make use of it where they deem it helpful.

## Initial views on options 1 and 2

- 20.33 Blocking and muting can work at scale because they enable all users to block or mute accounts that they consider to be producing illegal content of the kind identified above and as being particularly enabled by communications functionalities such as direct messages and commenting on content. We accept that this measure will not eliminate entirely the risk of users encountering those kinds of illegal content on a service (for example, a user may still encounter illegal content posted by an unblocked or unmuted account with which they are connected). Instead, we consider that this measure is more targeted and preventative, enabling users to reduce the risk of encountering illegal content from a particular account, or all non-connected accounts.
- 20.34 The evidence outlined above suggests that services and users alike consider blocking and muting tools to be an important and effective means of self-managing risk on U2U services. Below, we consider the potential costs and rights impacts of options 1 and 2.

---

<sup>653</sup> Twitter, 2020. [New conversation settings, coming to a Tweet near you](#). [accessed 29 August 2023].

<sup>654</sup> SNAP Inc, 2022. [Parent’s Guide: Snapchat’s Family Center](#). [accessed 29 August 2023].

<sup>655</sup> Meta, 2021. [Introducing new tools to protect our community from abuse](#). [accessed 12 September 2023].

<sup>656</sup> Meta, [Control who can send messages to your Messenger Chats list](#). [accessed 29 August 2023].

<sup>657</sup> Discord, 2022. [Blocking & Privacy Settings](#) [accessed 29 August 2023].

<sup>658</sup> The Behavioural Insights Team and Data Ethics and Innovation, 2021. [Active Online Choices](#) [accessed 12 September 2023].

## Costs and Risks

- 20.35 For both options 1 and 2, relevant services that are not currently implementing these options would incur direct one-off costs to make the system changes to enable muting and blocking functions, and there would also be ongoing costs of maintaining these changes.
- 20.36 The direct costs to implement these features are likely to be dependent on the complexity of the service's system, the nature of how users typically interact on a service and the extent of organisational overheads required to implement changes. These are likely to vary significantly across platforms and are likely to increase for larger platforms. In some circumstances, there may also be some cost synergies with the implementation of a measure, described earlier, which stops users from sending messages to non-connected users.<sup>659</sup>
- 20.37 We have estimated that this type of functionality for both options is likely to require a one-off cost in the region of 20 to 150 days of software engineering time, with potentially up to the same again in non-engineering time.<sup>660</sup> Making assumptions about labour costs, we would expect the one-off direct costs to be somewhere in the region of £9,000 to £140,000.<sup>661</sup>
- 20.38 In addition to the one-off direct costs, we expect this type of measure to require ongoing maintenance costs to ensure the functionality continues to operate as intended. We assume this would be 25% of the one-off costs and so we would expect it to be approximately £2,500 to £35,000 per year.<sup>662</sup>
- 20.39 We also recognise that global blocking of all non-connected users (with option 2) could fundamentally alter the community of a site. Users may be less likely to interact with other unknown users which could reduce engagement and use of a service. There may therefore be indirect costs from reduced revenue from lower use.
- 20.40 However, it is not necessarily always the case that use and revenue will fall. If users feel safer online, they may engage more with a service, albeit with a narrower set of other users. Without such measures, some users may leave a service entirely. The extensive availability and take-up of these measures across different types of services suggests that services think they add value for services and users. We regard the benefits that users accrue from blocking exceed the costs to them personally or they would not use the block functionality on services.
- 20.41 Interaction between user accounts differs across different U2U services, according to the functionalities that are employed. This means considerable variation in both the direct and indirect costs across different services.
- 20.42 The blocking of individual users is already widely implemented across larger U2U services, including across marketplaces (eBay),<sup>663</sup> pornographic services (PornHub),<sup>664</sup> payment providers (PayPal),<sup>665</sup> and file sharing (Google Drive).<sup>666</sup> In our view, this indicates that the

---

<sup>659</sup> This measure reduces the risk of harm from grooming. For more explanation, please see Chapter 18.

<sup>660</sup> For example, legal or project management costs.

<sup>661</sup> See Annex 14 for details of our assumptions on labour costs.

<sup>662</sup> See Annex 14.

<sup>663</sup> eBay. [Blocking a buyer on eBay](#). [accessed 12 September 2023].

<sup>664</sup> Pornhub. [How do I unblock a user](#). [accessed 10 September 2023].

<sup>665</sup> Paypal. [How Do I Block Another PayPal User](#). [accessed 12 September 2023].

<sup>666</sup> Google. [Block & unblock people in Google Drive](#). [accessed 12 September 2023].



costs of implementing the first option are likely to be reasonable for larger services, as the widespread use of blocking suggests that many larger online services consider any associated costs to be proportionate when compared to the value the feature provides to users. Individual user blocking is already widely offered across U2U services, while broader blocks on all non-connected users are less widely offered but we consider the costs that would be incurred by its implementation would not be significant.

## Rights and equality impacts

### Freedom of expression and freedom of association

20.43 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 or to 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 infringe on the rights of users.

### Privacy

20.44 We do not consider there to be any impact on the right to privacy, as we have not recommended as part of this measure that services extract or retain information relating to the application of blocking and muting, or accounts targeted by these features.

### Provisional conclusion

20.45 In summary, we consider that, as both a precautionary and reactive response to illegal content, user blocking and muting options 1 and 2 would be effective at preventing users from encountering certain kinds of illegal content, such as controlling and coercive behaviour; harassment, stalking, threats and abuse; hate; grooming; and encouraging or assisting suicide or serious self-harm. These user tools act as a safety net in the case of moderation tools failing to detect and remove illegal content.

20.46 While we believe the benefits are considerable, we are aware there would be costs from implementing options 1 and 2, including both the direct and indirect costs, and recognise that they could vary considerably across different types of services. Notwithstanding this uncertainty, our provisional view is that it would be proportionate to recommend that large services undertake the measures set out in options 1 and 2. There are a number of reasons for this:

- a) The offences these measures would tackle are all too common online and can cause significant harm. Our analysis suggests that the measures we are proposing would be effective in tackling them, thereby generating significant benefits. These benefits would be greatest where the measures are deployed on the largest services because they have the highest reach, meaning more users would have the tools to help protect themselves from illegal content.
- b) Given their scale, in general, it seems likely that the largest services would have the resources to bear the cost of measures of this nature. This view is borne out by the fact that many large services already have comparable processes in place.
- c) Individual user blocking is a reactive measure that users might employ to prevent users repeatedly sending them illegal content, but carries a residual risk from organised



harassment and the use of alternative accounts to continue sharing illegal content. We believe that residual risk can be addressed, at least in part, by services offering global blocking of all non-connected accounts, thereby offering a greater benefit to users.

- 20.47 By large service, we propose meaning a service with more than 7 million monthly UK users, as discussed Chapter 11 Overview of the Codes, paragraph 11.51.
- 20.48 Given the uncertainties about costs and the fact that there is limited evidence of smaller services undertaking these measures,<sup>667</sup> we are less confident at this stage that it would be proportionate to recommend that smaller services undertake these measures. We therefore do not propose to recommend these measures for services other than large services at this point.
- 20.49 We also outline above in paragraphs 20.4 to 20.11 how certain harms are most likely to manifest due to the existence of certain functionalities on a service. Therefore we propose to restrict the scope of this measure to large services which meet both of the following two conditions:
- a) **They have assessed themselves as being at risk of a relevant harm:** we propose to recommend these measures for large services that have assessed themselves as being at medium or high risk of any of the following kinds of harm: coercive and controlling behaviour; harassment, stalking, threats and abuse; hate; grooming; and encouraging or assisting suicide or serious self-harm.
  - b) **They have relevant functionalities:** we propose to recommend these measures for large services that meet **both** of the following:
    - i) Enabling users to interact by means of user profiles;<sup>668</sup> and
    - ii) Having at least one of the following functionalities: User connections; posting content; and user communication more generally (including but not limited to direct messaging and commenting on content).<sup>669</sup>
- 20.50 We are therefore proposing to recommend in our Code on CSEA offences and our other duties Code that large services which satisfy the criteria set out above should offer every registered user the option to:
- a) Block other user accounts. A user (the “blocking user”) should have the option to block each of:
    - i) An individual user, whether that user account is connected or unconnected to the blocking user; and
    - ii) All user accounts which are unconnected to the blocking user; and
  - b) Mute other user accounts. A user (the “muting user”) should have the option to mute specific user accounts, whether connected or unconnected to the muting user.

---

<sup>667</sup> In this context we define smaller services as any service which is not a large service.

<sup>668</sup> Includes information that is displayed to other users such as images, usernames, and biographies. Characterised by users creating a user profile that shows their identity.

<sup>669</sup> These terms are defined in the glossary at Annex [16].

## Measure to give users the ability to disable comments

---

### Harms or risks that the measure seeks to address

- 20.51 Above in this chapter, we have outlined that users may be at risk of encountering illegal content in the course of their engagement on U2U services. This includes encountering illegal content in the comments section which is embedded below shared content on many U2U services.
- 20.52 Some users may wish to share content without restricting it to specific users such as close friends or connections, whether to maximise the reach of the post or for other reasons. This leaves the user exposed to the risk of receiving illegal content from other users in the form of comments.
- 20.53 Unlike private messages, comments are visible to any users that have access to that content, either through being connected to the user who uploaded the content or by virtue of following a public account. The comments, and in particular any comments that contain illegal content, can therefore be accessed by other users, not just the content creator.
- 20.54 As outlined in Volume 2: Chapter 6E(Harassment, stalking threats and abuse offences), paragraphs 6E.69 and 6E.70; Chapter 6F Hate offences, paragraph 6F.49; Chapter 6D Encouraging or assisting suicide (or attempted suicide) or serious self-harm offences, paragraphs 6D.67-6D.69; Chapter 6C(CSEA grooming and CSAM), paragraph 6C.85-6C.87, and referred to above at paragraph 20.8, there is evidence that commenting on content is a functionality that may be used to spread illegal content and presents a particular risk of encountering certain priority offences such as harassment, stalking, threats and abuse; hate encouraging or assisting suicide or serious self-harm; and grooming.
- 20.55 Users may find themselves particularly vulnerable to illegal content in comments for a variety of reasons. For example, being a prominent public figure; being a target of a hate/harassment campaign; being the subject of a rumour; or being a member of a group with a protected characteristic (as discussed in more detail at paragraph 20.10 above).
- 20.56 For accounts with a large amount of followers, or content that is picked up by recommender systems and promoted widely, the number of people exposed to and potentially harmed by the comments can rise rapidly. This may be exploited by bad actors who use a piece of content as a vector to spread illegal content in the form of abusive hate speech, and threats of violence to the creator. For example, the amplification of illegal content enabled by the comments functionality is demonstrated in the following examples:
- a) As noted in paragraph 20.10c above, the example of the online abuse targeted at Premier League footballers indicates that users may send just one abusive comment to an individual,<sup>670</sup> but the targeted individual can receive comments from many users simultaneously<sup>671</sup> For those individuals, blocking all non-connected users may not be a preferable option given their public status but rather they may prefer not to receive hateful and abusive content in comments.
  - b) An investigation by The Washington Post into the experiences of seven female YouTube creators with large followings found that they had all been targeted with harassment,

---

<sup>670</sup> Ofcom, 2022. [Crossing the line: Seven in ten Premier League footballers face Twitter abuse](#). [accessed 7 September 2023].

<sup>671</sup> Ofcom, 2022. Pg 30

hate and misogyny through comments on their videos.<sup>672</sup> It highlighted that comments sections can become a meeting place for users with similar harmful intentions. This idea is supported by research conducted by Professor Matthew Williams and Mishcon de Reya which 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.<sup>673 674</sup> This research demonstrates that threatening, abusive or insulting comments may be shared, causing harassment or distress for the recipient.

- c) In the May 2020 series of attacks involving the Epilepsy Society’s page on X (formerly Twitter), perpetrators posted hundreds of flashing images and GIFs in replies (comments) in the comments section of posts, with the aim of triggering seizures in people with photosensitive epilepsy. Many reported seizures following the posts.<sup>675</sup>

## Options

- 20.57 To mitigate the potential harm produced by one user to another in the comments feature, we have considered the following possible measure: **Services that allow commenting on content should allow users to disable comments on content.** Users should be given an easy way to find and use a functionality to disable comments.
- 20.58 For the purposes of our discussion and recommendation, 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, visually accessible directly from the original content without navigating away from that content.
- 20.59 We note that some services offer users a range of comment control tools (see evidence on costs below). These are beyond the options we have carefully considered here. While we are supportive of these tools as a means of empowering users to exercise more control over comment functionalities, at this stage we have limited evidence around more granular controls, and have concerns given the risk of unintended consequences with regard to uneven impacts on freedom of expression and likely higher implementation costs.

## Effectiveness

- 20.60 Comment controls may protect users in different scenarios, for example:
- a) Users may pre-empt that a particular upload may lead to comments that contain illegal content, and choose to disable the comments to prevent them from having to engage with any comments at all. This may be because they have received illegal content in comments in the past, have a low risk appetite, or if the content they are uploading is particularly contentious and has the potential to spark abuse or harassment.
  - b) As well as protecting the content uploader, disabling comments would also protect other users (such as followers) who may encounter and potentially be harmed by the

---

<sup>672</sup> Lorenz, T., 2022. [YouTube remains rife with misogyny and harassment, creators say](#), Washington Post, 18 September. [accessed 7 September 2023].

<sup>673</sup> Williams, M., 2019. [Hatred Behind the Screens A Report on the Rise of Online Hate Speech](#), (p.26) Mishcon de Reya [accessed 30 August 2023].

<sup>674</sup> Williams, M (2021). *The Science of Hate*. Faber.

<sup>675</sup> Epilepsy Society, 2021, [Written evidence submitted by the Epilepsy Society \(OSB0008\)](#) (p.5-12) [accessed 30 August 2023].

illegal content when engaging with the user's uploaded content. This is particularly relevant for comments where the comment is an attack on a protected characteristic.

- 20.61 Additionally with regard to child users, by providing children the ability to limit interaction with their content in the form of comments, this option can support the reduction of non-consensually recorded (or 'capped')<sup>676</sup> CSAM or self-generated indecent images by reducing perpetrators' ability to interact with video and livestreaming content of children which, as outlined in paragraph 20.8c is used in the context of grooming or more broadly to incite and record sexual activity.
- 20.62 We do not have access to data to evidence how widely used comment disabling tools are by users of large U2U services, or how effective they consider this to be as a means of reducing the risk of exposure to illegal content. However, there are a number of cases whereby high profile figures have announced publicly their decision to disable the comments section on their uploads to social media after receiving abusive comments from other users.<sup>677</sup> Abusive behaviour which is likely to cause fear, alarm, harassment or distress may well generate priority illegal content under the Act. Moreover, notwithstanding limits on the empirical evidence base, as explained above, there are strong theoretical arguments which would suggest that the ability to disable comments would be an effective tool for reducing harm.
- 20.63 In Ofcom's 2022 Illegal Harms Call for Evidence, a number of stakeholders called on services to implement measures to allow users to have greater control over the comments section:
- a) 5Rights stated that users should be given the option of switching off comments as a way to prevent their exposure to harmful content.<sup>678</sup>
  - b) The Anti-Defamation League (ADL) recommended a variety of measures that services should adopt to allow users to control what they encounter, including restricting certain high-risk functionalities. These measures include comment filter settings and an option to hide comments.<sup>679</sup>
  - c) Refuge points to comment control measures as effective in helping prevent harm. It states that the ability to filter harmful comments and phrases is often used by survivors of tech abuse to prevent them from seeing triggering and abusive comments by the perpetrator, whether by using his own account or a fake account, or comments sent by his friends or family.<sup>680</sup>
- 20.64 Overall, we consider that an option to turn off comments is likely to be effective against the illegal harms we have identified in paragraphs 20.51-20.56 above.

## Costs and risks

- 20.65 Services that do not currently offer a functionality of this nature would incur one-off costs to make system changes and update the user interface. However, given this measure would only be adapting the ability to use an existing comments function, we would expect costs to be lower than introducing the blocking and muting features outlined above. We have

---

<sup>676</sup> Capping is defined as "the act of capturing imagery of videos of others performing sexual acts without their knowledge or consent." Source: InHope, 2021, [What is Capping?](#) [accessed 30 September 2023].

<sup>677</sup> Galluci, N., 2019, [Taylor Swift says turning off Instagram comments does wonders for self-esteem](#), *Masheable*, 6 March [accessed 12 September 2023]

<sup>678</sup> 5 Rights response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

<sup>679</sup> ADL response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

<sup>680</sup> Refuge response to 2022 Ofcom Call for Evidence: First phase of online safety regulation.

estimated that the direct costs of this measure would take approximately 5 to 50 days of software engineering time, with potentially up to the same again in non-engineering time. Making assumptions about labour costs, we would expect the one-off direct costs to be somewhere in the region of £2,000 to £50,000.<sup>681</sup>

- 20.66 In addition to the one-off direct costs, we expect this type of measure to require ongoing maintenance costs to ensure the functionality continues to operate as intended. We assume this would be 25% of the one-off costs and so we would expect it to be approximately £500 to £12,500 per year.<sup>682</sup>
- 20.67 In addition to the direct costs of implementing the function, there are also potential indirect costs that are more difficult to estimate. Indirect costs could arise if the measures lead to an increase in the disabling of comments and a decrease in user interaction with (non-harmful) content. This reduction in usage could in turn reduce revenue to services.
- 20.68 However, set against this, giving users the ability to disable comments may deliver some indirect benefits to services. We consider that for some users the inability to disable comments may result in them leaving the service which would impact the revenue of services, though the overall impact of these contrasting effects is difficult to estimate.
- 20.69 We also recognise that giving users the ability to disable comments under content could result in a negative impact when users are unable to reply to posted content. This is particularly relevant when it is defamatory or fraudulent content. For example, users would not have the ability to comment to explain that the content poster is engaging in a scam. This could result in instances of other users being defrauded, which might not have occurred if comments were allowed. Also if users started to disable comments on a widespread basis, the ability for other users to interact with content would be significantly reduced. Over time this could potentially lead to lower engagement and use of the service, or even users leaving a service altogether, with a consequential impact on service revenues.
- 20.70 We understand various large U2U services have already implemented measures to give users greater control of the comment functionality. This indicates that, even considering the costs outlined above, they consider it is an affordable and proportionate measure given the benefits it can provide to users. Examples of large U2U services that have implemented this type of measure are:
- a) **Instagram** enables users to disable all comments or block certain users from commenting.<sup>683</sup> It also allows for comment filters to be applied to filter certain words in from appearing in comments on posts.<sup>684</sup>
  - b) **X** allows users to restrict replies to Tweets by allowing only people the user follows or mentions to be able to comment.
  - c) **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’.<sup>685</sup> Users can also block comments from specified users, and filter comments, with the option to “hide offensive comments” or manually filter out key words. It also allows users to disable the comment functionality on Facebook Live videos

---

<sup>681</sup> For further details on our assumptions of labour costs, see Annex 14 paragraph A14.5.

<sup>682</sup> For further details, see Annex 14.

<sup>683</sup> Meta, 2021. [Introducing new tools to protect our community from abuse](#). [accessed 12 September 2023].

<sup>684</sup> Hootsuite (Hirose, A.), 2022. [How to Manage Instagram Comments](#). [accessed 12 September 2023].

<sup>685</sup> Meta, Facebook Help Centre. [Commenting](#). [accessed 12 September 2023].

for all users, or to restrict to followers, comments with over 100 characters, comments from accounts that are over two weeks old, and comments from accounts that have followed the content creator for at least 15 minutes.<sup>686</sup>

- d) **TikTok** allows users to disable comments on their videos, as well as setting rules around who can comment based on their connection. Settings for U16 users are set to ‘friends only’ for comments by default. It also has various comment filters including keyword filters.<sup>687</sup>
- e) **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.<sup>688</sup>

20.71 For smaller services, the evidence of existing services introducing comment disabling functionalities is less clear. Given that smaller services are likely to have lower revenues, the costs of introducing this measure could be a more significant burden to those services.

## Rights impacts

### Freedom of expression

20.72 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:

- a) be a choice made solely by the user concerned; and
- b) have no impact on the right of other users to express themselves freely on the service concerned in other ways.

20.73 For some users, if they do not have the ability to disable commenting, they may be discouraged from posting at all given the risk of harmful illegal content. This consideration is particularly acute for the public accounts of those with photosensitive epilepsy.

20.74 We therefore do not consider that it would amount to an infringement of any user’s right to freedom of expression and, by removing the risks for vulnerable users, the measure could in fact promote self-expression.

20.75 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, this measure if implemented as intended would not prevent users responding with their own content and linking to original content.

### Privacy

20.76 We do not consider there to be any impact on the right to privacy, as this option is not specifying that services extract or retain information relating to its application.

## Provisional conclusion

20.77 We have established that users are at risk of receiving illegal content from comment functionalities, as shown in the register of risk for: hate; harassment (including acts of

---

<sup>686</sup> Nerds Chalk, 2021. [How To Turn off Comments on Facebook Live](#). [accessed 12 September 2023].

<sup>687</sup> TikTok, [Commenting](#). [accessed September 12 2023]

<sup>688</sup> Sprout Social, 2022. [YouTube Comments: A Complete Guide](#). [accessed 12 September 2023]

epilepsy trolling), stalking, threats and abuse; grooming; and encouraging or assisting suicide or serious self-harm.

- 20.78 We consider that this proposed measure would be an effective means of reducing these risks, as it would give users the ability to proactively prevent other users from commenting on their uploads and potentially disseminating illegal content. It would also allow users to reactively disable the comments section after upload, if a post has attracted harmful comments. On services where a comments functionality does not exist, this measure would not be relevant. Given the prevalence and impact of the harms this measure would address, we provisionally consider that the benefits of applying it to large services would be significant.
- 20.79 There are some potential costs when implementing the proposed measure, however for large services we expect any direct costs to be manageable, and any indirect costs to be proportionate when considered against the expected reduction in risk of harm.
- 20.80 The fact that numerous large services are already implementing measures of this nature also suggests that our proposal is not likely to be overly onerous for these services. In view of this and in view of the potentially significant benefits the measure could confer, on balance we provisionally conclude that the proposal in question is proportionate for large services.
- 20.81 For smaller services, the lower number of users means the impact on harm reduction is likely to be lower. Given the likely lower benefits and uncertainty on whether the cost for small services may be towards the upper end of our estimate, it is not clear that it would be proportionate to apply this measure to all smaller services. Therefore, on balance we have decided not to propose this measure for smaller services as we are not confident at this stage that it would be proportionate.
- 20.82 In summary, we are therefore proposing to recommend, in our Code on CSEA offences and our other duties Code, that large services which enable users to comment on content and have assessed themselves as being at medium or high risk of at least one of the harms referred to above should offer every registered user the option of preventing any users from commenting on content posted by that user.

## Notable user and monetised verification schemes

---

### Harms that the measure seeks to address

- 20.83 As set out in the relevant chapters of the Register of Risk, impersonation for the purposes of fraud and foreign interference presents a risk to users.<sup>689</sup> This risk is heightened where bad actors take advantage of people trusting content or service design features they see online.
- 20.84 Online fraud causes a range of harms to individuals, principally financial loss, psychological harm and the compromise of important personal details provided under false pretences. The UK Government Fraud Strategy estimates that the total economic and social cost of fraud to individuals is £6.8 billion (2019/20).<sup>690</sup> Our Online Experiences Tracker shows that fraud is

---

<sup>689</sup> For further information, please see Volume 2: Chapter 6P (Foreign Interference Offence), Chapter 6O (Fraud and Financial Services Offences), and Chapter 6U (Governance, Systems and Processes).

<sup>690</sup> Home Office, 2023. [Fraud Strategy: Stopping Scams and Protecting the Public](#). [accessed 22 September 2023].



one of the most commonly experienced harms and one which the highest proportion of respondents are concerned about.<sup>691</sup>

- 20.85 Impersonation fraud can occur where bad actors trick users into engaging with fake online accounts of high-profile individuals, UK Government departments such as HMRC or the DWP, and financial institutions such as banks.<sup>692</sup> Impersonation of UK media outlets and think-tanks is also a tactic used by those engaging in forms of foreign interference.<sup>693</sup> Where users come to see these sources as credible, their engagement and sharing of content increases the risk of foreign interference.
- 20.86 The Register of Risk also describes how bad actors use impersonation tactics and fake identities to spread false information or commit romance fraud by luring victims into relationships or transferring money.<sup>694 695</sup> Online services with a large user base are particularly attractive to fraudsters as they make it easy for them to reach large numbers of people at low cost and with minimal effort. In addition, a large user base makes 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 re-posts.<sup>696</sup> There is significant evidence of influence operations occurring on social media services, video-sharing services, private messaging services, information-sharing services, and discussion forums and chat rooms.<sup>697</sup>
- 20.87 We see the impersonation of notable users and the broader use of fake accounts as being linked but separate issues. Considering the harms described above, we are interested in how high-profile users that may be subject to impersonation attempts are verified, labelled and featured in online verification schemes. We have looked closely at the user verification schemes and associated user profile labelling which some large online services provide. These schemes exist on a variety of online services, including services where users share content and their views, connect socially, or network for employment opportunities. We are also interested in how clearly the labelling of notable users is distinguished from the labelling of users who are verified for other reasons. Some of our evidence is presented in this section, and some is discussed in the “effectiveness” section. Given the evidence referred to above, we concentrate on the harms caused by impersonation of notable users, and we focus on large services that have assessed themselves as being at medium or high risk of fraud or foreign interference occurring on their service.
- 20.88 Our research about industry practice has focused on two types of user verification schemes which online services run. Throughout this chapter, we use the term ‘verification’ to reflect the language used by many services in describing what their schemes do, but note that services apply this term to a wide range of assurance methods. We use the word ‘labelling’

---

<sup>691</sup> This is set out in Volume 2: Chapter 6O (Fraud and Financial Services Offences).

<sup>692</sup> For more information on this, please see Volume 2: Chapter 6O (Fraud and Financial Services Offences), paragraphs 6O.10 and 6O.63.

<sup>693</sup> For more information on this, please see Volume 2: Chapter 6P (Foreign Interference Offence), paragraph 6P.50.

<sup>694</sup> For more information, please see Volume 2: Chapter 6P (Foreign Interference Offence), paragraphs 6P.47-6P.51.

<sup>695</sup> For more information on romance fraud, please see Volume 2: Chapter 6O (Fraud and Financial Services Offences).

<sup>696</sup> For more information, please see Volume 2: Chapter 6O (Fraud and Financial Services Offences), paragraphs 1.41-1.44, of the Register of Risk.

<sup>697</sup> For more information, please see Volume 2: Chapter 6P (Foreign Interference Offence), paragraphs 3.32-3.38.

to refer to the symbol and associated words added to a user’s profile which indicates they have been given verified status under a scheme and the reason given for that status. We recognise that verification can take many different forms, and in the ‘User Access’ chapter in Volume 4 we look more broadly at identity verification (‘IDV’) as a potential mitigation against the risks posed by anonymous user profiles. As noted above, we will also consult on requirements for optional identity verification schemes as part of the user empowerment duties for category 1 services in later phases of our work.<sup>698</sup>

20.89 The first of the user verification schemes we researched are what we call “notable user schemes” and the second are what we refer to as “monetised schemes”. In the first kind of scheme, services enable users meeting certain notability criteria, such as an account owner’s role in public life and the number of followers on the service, to take part in the scheme. Services add a label to the user’s profile to show that they were satisfied that their criteria were met, often confirming that the account is owned by, or operated on behalf of, the notable user. We note that these kinds of schemes have existed for over a decade, for example X established a tick or checkmark scheme in 2009 at a time when high profile users were complaining about people posing as them on the service.<sup>699</sup> Similar visible indicators have been used by other services and they have grown to be regarded as a visual cue of authority and sometimes as a status symbol.<sup>700</sup> Many large online services have presented them as a user support measure to help identify the “authentic” presence of notable users.<sup>701</sup>

20.90 More recently, a second kind of scheme has been introduced on some services such as Facebook, Instagram and X. Under these schemes, a much wider pool of users can be verified by paying money and meeting certain criteria which do not necessarily relate to notability. These schemes enable a user to have certain features such as a verification label, increased prominence, or additional features and functionalities compared to users not on the scheme. X introduced a monetised scheme initially known as “Twitter Blue” and later

---

<sup>698</sup> Further information on this can be found in paragraph 20.97 and in the chapter ‘Approach to Online Safety Regulation’.

<sup>699</sup> Ortutay, B., 2022. [Twitter's blue check: A history of the platform's verification system](#), *Fox 10 Phoenix*, 3 November. [accessed 24 August 2023].

<sup>700</sup> Waddoups, R., 2022. [The Complicated Legacy of Twitter’s Blue Checkmark](#), *Surface*, 15 November. [accessed 24 August 2023]; Stein, J., 2023. [The Life and Death of the Blue Check Mark](#), *Slate*, 29 April. [accessed 24 August 2023].

<sup>701</sup> For example, TikTok’s ‘How to tell if an account is verified’ page says it considers ‘a number of factors before granting a verified badge, such as whether the notable account is ‘authentic, unique, and active’. TikTok. [How to tell if an account is verified](#). [accessed 22 September 2023]; LinkedIn’s ‘Verifications on your LinkedIn profile’ page says that ‘Having verified information helps provide authenticity signals to others that you’re who you say you are.’ LinkedIn. [Verifications on your LinkedIn profile](#). [accessed 22 September 2023]; Snapchat’s ‘how to verify your public profile’ page explains that ‘Profiles will be verified upon meeting the below criteria: Authentic: Your profile must represent a real, registered business or entity. If we determine that information provided about the Profile is false or misleading, we will not issue or remove the verification and the account may become disabled. Notable: Your business or entity must be broadly known by the public.’ Snapchat. [How to verify your public profile](#). [accessed 22 September 2023].

renamed as “X Premium”.<sup>702</sup> Meta launched “a new subscription bundle” called “Meta Verified”.<sup>703</sup>

- 20.91 The schemes were discussed in the media as a new development in online services’ monetisation and revenue strategies.<sup>704</sup> Some media sources analysed the presentation of the new X and Meta schemes, commenting that accounts belonging to the original and new schemes appeared to be displayed with the same label.<sup>705 706 707</sup> This raised questions about whether users would be able to clearly distinguish between accounts that were verified under the different relevant criteria for notability or for a paid subscription.
- 20.92 For users to benefit from these verification and labelling schemes when engaging with content and accounts, they should be able to understand and see why another user is verified and what their verified status conveys about them. In the worst-case scenarios, confusion about verification schemes could cause harm to users by giving a sense of credibility and an amplified voice to a bad actor who is verified under a monetised scheme and is seeking to commit fraud or foreign interference. As a recent example, Martin Lewis, the Executive Chair of the UK’s biggest consumer help site, stated that a profile with a verified X Premium subscription checkmark was impersonating him to promote a cryptocurrency.<sup>708</sup> This meant there was a clone that could mislead users about financial advice and use the trust that other users place in both Martin Lewis’s reputation and in the verified status. Other publicised instances of the apparent misuse of verification schemes have attracted attention and confusion.<sup>709</sup>
- 20.93 We have examined our existing evidence about user awareness of labelling schemes from three studies and polls conducted in 2022 and 2023. They suggest that verified status is an important factor taken into account by users when deciding whether to engage with content or assessing if it appears to be genuine. It also highlights the importance of users being able

---

<sup>702</sup> 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. [About X Premium](#) and [About profile labels and checkmarks on X](#). [accessed 22 September 2023].

<sup>703</sup> 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. [Introducing Meta Verified](#). [accessed 22 September 2023].

<sup>704</sup> Ortutay, B., and The Associated Press, 2022. [How Elon Musk hopes to monetize Twitter’s ultimate status symbol: The blue checkmark](#). *Fortune*, 4 November. [accessed 24 August 2023]; McCallum S.M, and Gerken, T., 2023. [Facebook and Instagram paid verification starts in UK](#). *BBC News*, 16 May. [accessed 22 September 2023].

<sup>705</sup> 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].

<sup>706</sup> 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].

<sup>707</sup> 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].

<sup>708</sup> [Tweet](#) by Martin S Lewis of MoneySavingExpert on 3 April 2023. [accessed 22 September 2023].

<sup>709</sup> Sardarizadeh, S., 2022. [Twitter chaos after wave of blue tick impersonations](#). *BBC News*, 12 November. [accessed 24 August 2023].

to see why a label is placed on a profile and to be able to understand what it signifies. Both our adults', and children and parents' media use and attitudes trackers look at the experience of UK users online and their attitudes towards this. Participants were asked to judge whether a social media post appeared to be genuine and why they came to their conclusion. Half (51%) of the adult social media users who correctly identified the genuine content stated that the "verified tick" label was an indicator of the post's validity.<sup>710</sup> By way of comparison, among the 80% of respondents aged 12-17 who correctly identified an NHS Instagram post as genuine, nearly three in ten (28%) said the presence of a verified tick was an indicator of credibility.<sup>711</sup>

20.94 Another survey also gives some insight into whether users look out for verification labelling when using social media. In a poll conducted by YouGov for Ofcom, nearly three in ten (28%) UK internet users aged 16+ claimed that they use verification labels when deciding to follow or interact with an account on social media.<sup>712</sup>

20.95 In summary, the evidence discussed in this section shows that some forms of impersonation are a tactic used by those seeking to commit offences online. While this could take place within a wide range of contexts and online services, as explained above our evidence suggests that the higher number of users on large services means they are more attractive to those seeking to commit fraud and foreign interference. We consider that the corresponding risk of harm may be amplified where users are confused or unclear about labelling schemes they see as indicators of trust on large services. We therefore decided to consider a measure that seeks to set expectations for large services to operate their verification schemes in ways that can help address risks of fraud and foreign interference through the impersonation of notable users.

## Options

20.96 We considered three options to address the harms of impersonation on online services and the role of schemes which verify or include notable users:

- a) **Option A:** Rely on the user empowerment and user identity verification duties for Category 1 services created by the Act to address our concerns;
- b) **Option B:** Propose that large online services should establish and maintain a notable user verification system that meets certain criteria; or

---

<sup>710</sup> Ofcom, 2023. [Adults' Media Use and Attitudes report 2023](#). 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 44% of adult social media users who correctly identified a Money Saving Expert Facebook post as genuine, 51% identified the verification tick as amongst their reasons for making this judgement.

<sup>711</sup> Respondents aged 12-17 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. Of the 80% of who correctly recognised that it was a genuine NHS post, nearly three in ten identified the inclusion of a verification tick as one of the factors behind this judgment. Source: Ofcom, 2023. [Children and Parents: Media Use and Attitudes](#). [accessed 21 September 2023].

<sup>712</sup> 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%) use verification labels when deciding to follow or interact with an account on social media. A further fifth (22%) said they use 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](#) via YouGov panel. [accessed 21 September 2023].

- c) **Option C:** Retain these large services' flexibility to choose whether to operate a verification scheme and set out criteria that relevant services over a certain userbase threshold should follow, where they choose to operate such a scheme.

## Effectiveness

- 20.97 Option A: We carefully considered the relevance of section 64 of the Act, which will require a provider of a Category 1 service to offer all adult users an option to verify their identity, and section 15(9), which will allow users to filter out non-verified users. Whilst we consider these functionalities may be relevant to the risks discussed above, the verification required by these user empowerment and user identity verification provisions is not intended to replace existing notable user schemes and monetised schemes, nor affect whether services introduce new versions of such schemes. We therefore expect services to continue operating these other types of schemes after these provisions come into force. As such, we believe it is warranted to consider a separate Codes measure aimed at addressing the specific risks of harm arising from notable user impersonation, as set out above and in the Register of Risk.
- 20.98 Option B: We considered and do not currently favour the option of placing an expectation on services to operate verification schemes for notable users. We consider option B would be materially more intrusive than option C, as it would require services to implement a scheme where they do not already operate one, and we consider that the evidence is not strong enough at this stage to justify this additional step.
- 20.99 Option C: This option would recommend that a service that chooses to operate a notable user verification scheme (whether free to join or subscription-based) should do so against two criteria we explain below. This option would also apply these criteria to any monetised user verification schemes that a service operates. By doing this, services would have clear internal policies for operating their schemes, and provide to users a clear explanation about schemes which label some users as having a particular status, setting them apart from other users. Consequently, other users would be better prepared to understand the basis for that label being applied when they are deciding whether to engage with content posted by a labelled user.
- 20.100 We have evidence of the harm caused by the impersonation of notable and high-profile users and of the role of user verification schemes that some online services operate. We have reviewed publicly available information about user verification schemes on a range of services including: Facebook, Instagram, X, TikTok, LinkedIn, Snapchat, YouTube and Pinterest.
- 20.101 This option is based on balancing the risk of harm to users from illegal content and the evidence we have gathered to date of impersonation and industry solutions to help determine user authenticity, alongside the importance of services being able to make their own business decisions and generate revenue. It also rests on the premise that online providers partly choose to operate these schemes to build a sense of credibility for their services and for verified users.
- 20.102 As discussed above, our initial evidence base suggests that verified status is an important factor taken into account by users when deciding whether to engage with content or assessing if it appears to be genuine. We carried out a literature review to establish whether there is an evidence base for the claim that verification schemes are important for establishing user trust and content credibility and have a meaningful effect on user

behaviour. The results from our literature review present a varied picture.<sup>713</sup> This is in part due to academic interest in how other factors interact with verification 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 in 2009.

- 20.103 In general, the literature supports the fact that users are aware of verification schemes and that account credibility is a factor in increasing user trust.<sup>714</sup> However, the literature specifically linking credibility to verification or account authenticity is limited.<sup>715</sup> We conclude from this literature review that verification and account authenticity are likely to increase trust and credibility. However, they interact with a variety of other factors such as knowledge of the verification scheme and celebrity or congruity (e.g., when influencers promote material that does not ‘fit’ with user perceptions of the account holder).
- 20.104 Our approach to designing this option is based on the aim that services design, improve, and operate their schemes based on their own assessment of the risk of harm arising from impersonation on their service. This option also aims to recognise diversity in the design of verification schemes. It should bring greater clarity and understanding about all kinds of verification schemes, and also set additional expectations for services to responsibly operate their schemes that are specifically intended to label notable users.
- 20.105 We include below an overview of the rationale behind option C and describe two key criteria we consider would help to ensure relevant schemes are effective in preventing harm. These are expanded on in the following two sections.
- a) First, relevant services should have clear internal policies on the operation of notable user verification schemes and monetised schemes.
  - b) Second, relevant services should improve public transparency for users about what verified status means under notable user verification schemes and monetised schemes.
- 20.106 This option does not entail that schemes should be identical across different services. We recognise that it is desirable for them to operate their schemes in the most appropriate way for the purposes and features of their service and their userbase. For example, it is for individual services to decide whether they want to verify accounts posing as real people for the purposes of parody, fan accounts, or satire.

## Clear internal policies on the operation of verification schemes

- 20.107 Here we explain why we consider that services operating notable user or monetised schemes should have clear internal policies about which users are eligible to be verified and

---

<sup>713</sup> 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]; 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]; 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].

<sup>714</sup> Vaidya, T. et al. 2019; 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]; 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].

<sup>715</sup> Kapitan, S., and Silvera, D., 2016.



how the schemes are operated. We set out below what we think the features of an internal policy and associated processes should include. User-facing information should be representative of this internal documentation, to ensure consistency between user understanding of schemes and how they are operated in practice.

- 20.108 Eligibility and checks performed: We would expect the internal policy to set out how and why verified status may be granted under each relevant scheme that a service is operating. This would include information such as the criteria for verification under each scheme, and the reasons why verified status might be rejected or removed. For notable user schemes, the policy should set out how a service would satisfy itself that an account verified under its notable user scheme is operated by or on behalf of the notable person or organisation that the account is held out as being operated by or on behalf of. It should also set out the steps that a service takes to satisfy itself that the notable user holds the position or role they claim to hold, for example, an elected political figure. A policy that describes which user attributes a particular scheme seeks to verify and how it does this would help ensure consistency in staff decision-making and operation of the scheme.<sup>716</sup>
- 20.109 Safeguards against misuse: The policy should also set out safeguards that the service takes to ensure that user profile information, such as username and bio, is not modified without the relevant service reviewing and consenting to that change. This should help avoid misuse or abuse of the scheme going undetected, for example a user being verified and then changing their details so as to hold themselves out as a different person, and relying on their verification label in order to carry out impersonation. We consider a service's policies should also set out the circumstances and frequency of reviews to confirm whether the user profiles of relevant users continue to qualify to be labelled. Ensuring policies cover these matters should help to ensure that users with verified labels continue to represent the users that a service originally verified.
- 20.110 Communication to staff: To help consistency between policy application and user information, the internal policy should be in writing and easily understandable for staff implementing it. Any changes or updates should be actively communicated to staff. This communication should be aimed at staff whose role is directly and indirectly related to operation of schemes. This is because some services use these schemes to determine how users are treated or prioritised on a service. This should reduce the risk that an account or its content are given privileged status by one part of the service based on misunderstanding or a lack of consideration about what verification means or the processes underlying it.
- 20.111 Design of the scheme to mitigate harms: We expect services to actively consider how the design of their notable user verification schemes and monetised schemes can reduce the risk of harm to users. As part of this, services should consider what steps they can take to mitigate the risk of their schemes being used for fraud and foreign interference in the context of the service and its userbase.<sup>717</sup> Relevant staff at online services should have a full understanding of the role that well-run verification schemes could play in tackling online harms like fraud and foreign interference. This should guide staff when considering decisions about reviewing and improving operation of schemes, which should occur at regular

---

<sup>716</sup> For example, UK Government guidance sets out examples of how to decide how to check someone's identity. Source: Cabinet Office and Government Digital Service, 2023. [How to prove and verify someone's identity](#). [accessed 21 September 2023].

<sup>717</sup> For example, the risks may be different between social or professional networking sites, or if notable users such as elected political figures or major financial brands use the service to communicate with other users.



intervals. As part of these reviews, service providers should, if they consider it appropriate, have regard to user feedback and reporting, user experience testing, and/or engagement with persons with relevant expertise. It is up to services to choose whether and how to use these to inform their reviews, but our view is that they would provide useful sources of information in ensuring that schemes are fit for purpose and to mitigate the risk of relevant offences being committed or facilitated on the service.

## Public transparency for users about what verified status means

- 20.112 Here we explain why we consider a service should make it clear to their users why verified status has been granted and the steps it has taken to do this. Providing visible labels and clear explanations would help users understand what the verification process and the associated labelling on the user's profile actually represent. These actions would help inform users who take verification labelling into consideration when making decisions about illegal content which could cause them harm.
- 20.113 Visible and well-explained verification labels: Where verification labelling appears on a user profile, it should be straightforward for users to find out why verified status has been granted and the user profile labelled in that way. We have observed on services that verification is often denoted by a badge, tick or checkmark symbol which is generally prominent on a verified user's profile, and may also be visible where users see and engage with content posted by the verified user. It is likely that these labels are the most widely seen information about verification. Sometimes services provide brief additional information explaining the label, such as via a "right click" or a "hover text", but we have observed that this is an area for improvement across services. It should help users to make informed decisions about whether to engage with content if they can easily access information about why verified status has been granted. As noted above, our research suggests that a "verified tick" is a factor that users consider when deciding whether to engage with other users' content.
- 20.114 Publishing user-facing explanations and FAQs: Providing users with a more detailed explanation of a service's schemes is also important and this should be clear and accessible for users to find and read. The explanation should be consistent with the service's internal policies, including setting out why and how profiles are labelled or why a label may be removed. There is currently considerable variation in the depth of explanations provided by services to users about whether and how services decide if a profile is the authentic presence of a user. We have observed that explanations provided for users often do not convey if different or additional kinds of checks are made on accounts that are presented as belonging to notable users with a particular role or position in public life or that have a particular status within the service itself (for example, reaching a threshold for followers or paying a subscription).
- 20.115 Providing more detailed but nevertheless accessible user-facing explanations can support media literacy (discussed in more detail below) and promote better understanding of service design features. Whilst we recognise that many users may not read the information, it may be summarised and cascaded by other users who have an interest in analysing these features. We also recognise that in some areas, services may wish user-facing information only to present a high-level summary of a service's internal verification policy in order to mitigate any risk of bad actors using the public information to circumvent or abuse certain aspects of a scheme.

- 20.116 Providing clarity about differences between schemes: As discussed earlier, the landscape of verification schemes offered by online services is changing rapidly, such that clarity about the different schemes a service operates will be particularly useful. As discussed above, some services operate only notable user schemes while others operate multiple schemes aimed at different sections of their userbase or as part of monetisation or other strategies. We believe that there is a risk of confusion on the part of users as to why a user is verified under a particular scheme where users may have difficulty distinguishing between the schemes and their associated labelling. In a worst-case scenario, misunderstandings could be exploited by subscribers to monetised schemes and lead users who are influenced by verified status to engage with content that is intended to deceive for fraudulent purposes or covert influence. Our provisional view is that this risk could be mitigated by services being clear in their user-facing communications about the differences between schemes and what scheme a user is verified under. If users cannot distinguish between multiple verification schemes on a service, this could have a longer-term consequence on user media literacy or their trust in content they see on services.
- 20.117 Media literacy: Greater awareness of user verification schemes may help users' ability to distinguish between genuine and fake accounts. As set out earlier, our evidence shows that at least around a third of surveyed users take these symbols into account when choosing whether to engage with content. Increasing awareness of verification schemes should enable more users to make better informed choices about what content to engage with and reduce the risk of them experiencing harm as a result of engaging with fraudulent and foreign interference content posted by bad actors. Published explanations of the schemes may also help notable users to decide whether to become part of the schemes.<sup>718</sup> As noted above, the option specifies that users should be able to easily access information about why a user is verified from that user's profile. As such, the greater the number of notable users who are part of such schemes, the more likely it is that other users will come across the profile of a notable user and therefore be able to learn about a service's scheme. In theory, the more notable users that are verified under notable user schemes, the greater the possibility that other users can identify the authentic presence of notable users, thus lowering the risk of users falling victim to bad actors carrying out impersonation.

## Costs and Risks

- 20.118 We have considered the costs of operating and improving verification schemes alongside our evidence of user awareness and efficacy of verification schemes. If this option was applied to services that did not operate a verification scheme, it would include the costs of setting up and maintaining a scheme, which could be material. If it was only applied to services that choose to operate a verification scheme, the costs associated with the measure would not include the costs of setting up and operating the verification scheme. Costs which a service with such a scheme will incur in any case. Nevertheless, we accept, there could be additional costs of implementing this measure for any service that does not currently operate its verification scheme in line with the two criteria noted above.
- 20.119 The first part of this measure would place an expectation on relevant services to have clear internal policies on the operation of notable user verification schemes and monetised

---

<sup>718</sup> For example, Snapchat and other services publish information pages to explain their verification schemes to business and notable users. Source: SNAP Inc. [How to verify your public profile](#). [accessed 5 September 2023].

schemes. The additional costs could include developing or improving the appropriate internal policies and processes and training staff to apply them. The policies would need to cover which users are eligible for verified status and how services will apply their verification policies. Staff working on this service feature would need to have a good understanding of these policies. Some services may also have costs associated with implementing an updated policy for their specific scheme or schemes(s) if their updates are more rigorous than their existing policies. In other cases, services may decide they can follow the measure by creating more detailed documentation than they currently use and using this as a basis to improve public explanations of their schemes.

- 20.120 Some services may incur additional costs associated with designing their schemes to mitigate harms or with any safeguards they put in place, such as reviews for users they are labelling as notable individuals or organisations. Services may also choose to deploy additional measures to their schemes, such as a means of checking whether the identity of a verified account is being misused (for example, a verified account changing its name to appear the same as another user).
- 20.121 Where services consider that their verification scheme(s) could be better designed to decrease the risk of harm to users (for example to reduce the risk that users misunderstand the schemes), then services would need to consider how they can redesign the scheme(s). This could represent a substantial cost if the scheme(s) needed to change materially. While the policies and training associated with this measure may be more than some services have now, we anticipate that all services 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 would tend to reduce the total additional cost.
- 20.122 The second part of this measure would place an expectation on services to improve public transparency for users about what verified status means. Services may incur service design and engineering costs to make any necessary system changes, such as making the description of the verification scheme easily accessible to users. For some users that join a monetised scheme, the appeal of verification could be that its label is similar to those appearing on notable users' profiles. Our measure sets an expectation that services take steps to ensure the differences between the schemes are clearly communicated to users. This could disincentivise some users from joining them, reducing services' revenue. However, while this could be a disincentive for some, all of the monetised schemes we have examined state that there are other benefits to applying, such as exclusive features, such that ensuring users understand the difference between the schemes may have only a limited impact on take-up.
- 20.123 Alongside costs we have also considered benefits to both services and users. There could be an indirect benefit to services from this measure through the potential network effect in trust. Where some users see value in joining schemes, or using services with well-operated schemes, it could drive economic benefits to a service.
- 20.124 The option would therefore have some costs which are likely to vary by service. The scale of the cost would depend on the level of changes, if any, needed by an individual service to act in accordance with the measure. Partly because 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 services may only need to make small adjustments to the operation and/or communication of their schemes, and some services may need to make more significant changes.

## Rights impacts

### Freedom of expression and privacy

- 20.125 We expect this measure would have no impact on users' freedom of expression. All of the verification schemes currently operated by in-scope services are voluntary and a user's access to a service is not dependent on being verified.
- 20.126 We recognise that services may require users to provide personal information in order to be labelled under a relevant scheme. Given that verification is voluntary and that services would have to comply with applicable privacy and data protection law when collecting and processing users' data, we consider that any potential infringement of privacy should be limited.

### Provisional conclusion

- 20.127 As set out above, impersonation fraud online is a significant and growing problem, which imposes substantial costs on individuals and society. We believe that well-run notable user verification schemes can play a valuable role in enabling users to make informed decisions about the content they choose to interact with, helping them to identify potentially illegal content. Inversely, poorly operated and communicated schemes could introduce more risks than benefits for users who place trust in them.
- 20.128 We are proposing that services should have, and consistently apply, clear internal policies for operating notable user verification and paid-for user verification schemes and improve public transparency for users about what verified status means in practice. We consider that key risks for the schemes, as outlined above, are users falling victim to fraud or foreign interference. As such, we are proposing that this measure applies only to services that have assessed themselves as being at medium or high risk of fraud or foreign interference. Given the greater risk of these harms arising through impersonation on services with a larger user base, we are also proposing that the measure apply only to large services.
- 20.129 While costs are likely to vary by services and there is uncertainty on the precise level of costs, our provisional view is that it is likely to be 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 user verification schemes. The scale of the challenge posed by impersonation to commit fraud or foreign interference, and hence the potential benefits of the measure, are considerable on such services and are likely to justify the costs associated with them. Large services are likely to have the resources to bear any costs of this measure. We recognise these services could choose to discontinue their verification schemes if they find the costs of the measures excessive, but consider that providers are unlikely to remove a longstanding element of their services which some users appear to value. Given the benefits are likely to be smaller for services with lower reach and there is uncertainty on the precise level of costs, it is not clear the measure would be proportionate for smaller services. For all these reasons, we are not proposing to recommend this measure for smaller services at this time.
- 20.130 We therefore propose to recommend in our other duties Code that relevant services should have and apply consistently internal documented policies regarding the operation of notable user and monetised user schemes, and should provide certain information on the user profile of a labelled account and a user-facing description of the scheme in question. Our

view is that the proposed measure would assist services in complying with the illegal content safety duty in section 10(2) of the Act.

20.131 We recognise that impersonation is a factor in a much broader range of harms such as romance fraud, fraud on online marketplaces and in the sharing of disinformation online. However, this proposed measure is targeted to address a particular way in which impersonation manifests, and at this stage, we have focused on remedies for impersonation of notable users for the reasons described above. We note that other types of services (such as marketplaces or dating services) also operate forms of verification schemes and we would like to expand our policy thinking as our evidence base on illegal harms and remedies grows, and as Ofcom consults on the user empowerment duties.<sup>719 720</sup> The themes covered by this proposal are relevant to the optional identity scheme duties in the Act. We will propose guidance in respect of those duties for Category 1 services in later phases of our work.

---

<sup>719</sup> Pets4Homes has introduced ID verification for transactions between sellers and pet purchasers. It describes the process and benefits for verified users on its [“Why are we introducing ID verification?”](#) page. [accessed 22 September 2023].

<sup>720</sup> The dating service Tinder operates an optional process that allows users to show that the photos on their profile look like them. It describes [“How Does Photo Verification Work”](#) on its service. [accessed 22 September 2023].

# 21. User access

## What is this chapter about?

This chapter considers whether blocking users who have posted the most harmful types of content from using a service could play a role in improving online safety. Ofcom recognises the considerable implications that recommendations we make around users' ability to access a service could have on user rights and have carefully considered this in developing our proposals.

## What are we proposing?

We are making the following proposal for all U2U services:

- **Services 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').**

We are also planning further work on a measure, potentially for all U2U services:

- **Services should block the accounts of users that share CSAM.** We are gathering more evidence to inform the detail of any such measure. We aim to consult on the full detail of such a measure next year.

## Why are we proposing this?

There is some evidence that blocking users who post the most harmful types of content from accessing a service can help combat online harms. However, we have provisionally decided to proceed cautiously in this area given the significant implications restricting users' access to the internet would have for fundamental rights such as freedom of speech, and the fact that there are gaps in our evidence base about technical options for blocking users. We therefore focus the proposals in this chapter on a small number of the most serious types of illegal harm.

Given our current evidence base, we believe it is proportionate to recommend measures requiring the removal of proscribed organisations because taking any intentional action for the benefit of a proscribed organisation is an offence. Removing proscribed organisations' accounts should make it more difficult for these organisations to communicate online.

Provisionally, we consider that a measure recommending that users that share CSAM have their accounts blocked may be proportionate, given the severity of the harm. We need to do more work to develop the detail of any such measure and therefore aim to consult on it next year.

## What input do we want from stakeholders?

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

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

- What are the options available to block and prevent a user from returning to a service (e.g. blocking by username, email or IP address, or a combination of factors)? What are the advantages and disadvantages of the different options, including any potential impact on other users?

- How long should a user be blocked for sharing known CSAM, and should the period vary depending on the nature of the offence committed?
- There is a risk that lawful content is erroneously classified as CSAM by automated systems, which may impact on the rights of law-abiding users. What steps can services take to manage this risk? For example, are there alternative options to immediate blocking (such as a strikes system) that might help mitigate some of the risks and impacts on user rights?

## Introduction

---

- 21.1 User access concerns a user’s entry on to a service and ability to use the functionalities present on that service. It covers not only initial access, like at first sign up, but also a service controlling access throughout a user’s journey, including measures taken in response to identified illegal behaviour. We see mitigations concerning user access to be related exclusively to user-to-user (U2U) services and not search services, given that users are not required to hold accounts to make use of search services, or use these accounts to share content in the same manner as U2U services.<sup>721</sup>
- 21.2 Under their illegal content safety duties in the Act, regulated U2U services must take certain steps to reduce the risk of harm to users from illegal content 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)).
- 21.3 Effective user access measures can prevent illegal content from appearing and spreading on services and reduce the risk of repeat offending. User access measures are related to services’ content moderation processes, as they can be used as sanctions in response to upheld complaints. Terms of service and complaints processes play an important role in ensuring that users know about the procedures around these measures and have appropriate information and potential redress where they are applied.
- 21.4 Services have a range of possible tools to control and understand who is accessing their service or parts of their services, and to remove access from users, including identity assurance and verification, blocking, and limiting functionality access. We consider these when assessing our recommendations in this area. We focused our assessment on two high-level possible user access recommendations.
- 21.5 First, we consider **blocking or suspending users’ access to services in response to illegal behaviour** from (paragraph 21.8). Ofcom recognises the implications that recommendations we make around user access could have on user rights (from paragraph 21.38). As such, we are proposing to recommend two measures in respect of blocking or suspending users’ access to services:
- a) One recommendation to address proscribed organisations’ access to services. We provisionally consider that it is proportionate based on current evidence on effectiveness, costs, and user rights implications to recommend that services should remove proscribed organisations’ access to in-scope services. We have drafted an accompanying Code measure for consultation.

---

<sup>721</sup> Mitigations concerning user access will also be relevant to services hosting provider pornographic content as defined in Part 5 on the OSB.



- b) We are committed to protecting children online from the worst kinds of illegal harm. We discuss below the difficulties in making a detailed recommendation at this stage, but we will aim to explore a recommendation early next year that services should block users who post CSAM content. However, we need to do more work on the detail of such a measure and therefore, through this consultation, we are requesting further evidence around the scenarios in which users who post CSAM content should have their access to services restricted.
- 21.6 We assess the impact of these measures below and explain our provisional view that they would be proportionate interventions.
- 21.7 We then also consider **verifying users' identity** (from paragraph 21.73) or **specifically their age** (from paragraph 21.96) as a mitigation for illegal harms. Our intention is to issue further guidance on age verification in the coming months. While we do not propose to recommend identity verification in our Codes for illegal harms, we note that, under the Act, 'Category 1 services' have an additional specific duty to offer optional identity verification as a user empowerment tool. We will issue guidance in respect of the user empowerment duty for Category 1 services in later phases of our work.

## Blocking users' access to services following instances of illegal activity

---

### Harms that the measure seeks to address

- 21.8 For certain severe kinds of illegal harms, after content take down, risk may be presented by the offending user's continued access to the service. This is because in many cases these users repeatedly and persistently post illegal content or engage in illegal contact online. Our Register of Risks entry on Intimate Image Abuse points to continued and escalating illegal behaviour.<sup>722</sup>
- 21.9 Other evidence also points to the fact that many people who post illegal content do so repeatedly:
- a) A **Meta report into intent of CSAM sharers showed "patterns of persistent, conscious engagement with CSAM and other minor-sexualising content if it existed" when 200 accounts that were reported to the National Center for Missing & Exploited Children were analysed.**<sup>723</sup> Similarly, one research study found that, of a group of 78 perpetrators of child sexual abuse, 42% had attempted to collect all images in an abuse series, or of an individual, indicating a likelihood of persistent offending.<sup>724</sup>
  - b) Refuge's survey of 2,264 adults and interviews with 18 survivors of domestic abuse found that 48% of the female survivors of harassment and abuse who responded said that the abuse they experienced from the perpetrator on social media got worse over time.<sup>725</sup>

---

<sup>722</sup> Please see Our Register Of Risks Volume 2: Chapter 6M (Intimate Image Abuse offences), paragraph 50.

<sup>723</sup> Meta, 2021. [Understanding the intentions of Child Sexual Abuse Material \(CSAM\) sharers](#). [accessed 6 June 2023]. Subsequent references throughout.

<sup>724</sup> Steel, C.M.S., Newman, E., O'Rourke, S., Quayle, E., 2021. [Collecting and viewing behaviors of child sexual exploitation material offenders](#), *Child Abuse and Neglect*, 118.

<sup>725</sup> Refuge, (Eagleton, J.), 2021. [Unsocial Spaces](#) page 22.

- c) Fraudsters can target people multiple times over a period of time, with one case study showing a person who conned 80 victims out of over £400,000 between 2005-2021. One of these victims was defrauded over a period of 14 years.<sup>726</sup>
  - d) J.M. Berger and Jonathan Morgan’s report, “The ISIS Twitter Census”, refers to internal documents from the terrorist group that highlighted the importance of remaining an influential presence on social media to continue spreading the ISIS message.<sup>727</sup>
  - e) TikTok shared in a recent enforcement update that 90% of repeat violators violate using the same feature consistently, and over 75% violate the same policy category repeatedly.<sup>728</sup>
- 21.10 This demonstrates that continued access for users who commit some illegal harms poses a high risk for services of those illegal harms being repeated.

## Blocks and strikes definitions

- 21.11 Many services have addressed these harms and the problem of repeated violative behaviour by establishing systems of access-based sanctions for users who breach their terms and conditions.
- 21.12 Services refer to ‘strikes’ to describe a record of a user or account having contravened a service’s policies in some way. These can come with a warning to the user about their behaviour and be paired with temporary or permanent limitation of access to certain functionalities like content upload, streaming or contact with other users. Strikes are an escalating system, with more strikes meaning users move through and toward more severe enforcement action. Strikes offer the opportunity for users who have unintentionally behaved in harmful or illegal ways to amend their behaviour, which reduces repeat offending by some users.
- 21.13 A service ‘blocking’ users involves removing them from a service and often taking steps to prevent them from returning, although some services may choose to remove an account outright. It is usually deployed for serious or multiple infringements of service policies. A user can be temporarily blocked or have their account and access permanently blocked (often referred to as a ‘ban’). Chapter 20 on Enhanced User Controls deals with and makes recommendations regarding U2U blocking.
- 21.14 There are several different technical ways to implement user blocking, including by username, contact information, device or network identifiers, or a combination of these methods and others. We are also aware that some services monitor newly created accounts for behaviour patterns matching that of previously blocked accounts to prevent the return of offending users.

## Options

- 21.15 In assessing potential recommendations for our Codes which address the harms caused by users repeating illegal behaviour, we consider recommending the broad application of a

---

<sup>726</sup> HM Government, 2023. [Fraud Strategy: Stopping Scams and Protecting the Public](#), page 25.

<sup>727</sup> Berger, J.M. and Morgan, J., 2015. [The ISIS Twitter Census](#), page 55.

<sup>728</sup> TikTok, 2023. [Supporting creators with an updated account enforcement system](#). [accessed 6 June 2023]. Subsequent references throughout.

strikes and blocking system. We also consider whether immediate blocking from a service would be proportionate, examining harms where risk of repeat behaviour is high.

21.16 In this area, we consider the following options for Codes measures:

- a) 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;
- b) 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 CSAM; or
  - ii) Remove accounts run by or on behalf of proscribed organisations.

21.17 It is already a requirement of the Act for services to remove illegal content where they are aware of it. Through these options, we are considering recommendations around other means of preventing users from sharing illegal content or behaving illegally, in relation to blocking users' access to the service where they are found to have shared illegal content or engaged in severe illegal behaviours.

21.18 We consider these recommendations specifically in the context of services' duties in relation to illegal content and commission or facilitation under the Act. However, we acknowledge 'strikes' and 'blocking' in a broad sense can encompass any actions that services take against user accounts found to be in contravention of services' terms of use. Some of our evidence base below includes such actions.

## Current strikes and blocking practices

21.19 We currently have evidence to inform a high-level understanding of current strikes and blocking policies being deployed by services. However, our evidence is more limited on how these systems operate in practice, and their effect on the availability of harmful content and users who share harmful content.

### Strikes

21.20 Much of our evidence shows that, where strike systems are employed by services, the behaviour which can lead to a strike against an account is broad, and the system of escalation varies.

21.21 Many services have strike systems in place, including YouTube, Meta services, and TikTok. Some take a contextual approach. Meta says that "a strike depends on the severity of the content, the context in which it was shared and when it was posted".<sup>729</sup> Others have a more cumulative method, viewing strikes as a fixed points system. **UK Babe Network's system works through users accumulating strikes, "similar to penalty points on a driving licence."**<sup>730</sup>

21.22 Evidence also shows some services view sanctions – warnings or punishments given to a user for behaviour against policies – as like a strike; if the user continues their behaviour after

---

<sup>729</sup> Meta, 2022. [Counting strikes](#). [accessed 6 June 2023]. Subsequent references throughout.

<sup>730</sup> UK Babe Network response to 2022 Illegal Harms Call for Evidence.

said sanction, they may face further restrictions to their access to the service or certain functionalities. **For example, EA uses a warning before banning users which it finds effective, stating that “85% of players who receive feedback about their behavior under the Positive Play charter changed their behavior. We didn’t have to ban them.”**<sup>731</sup>

## Blocking

- 21.23 Our evidence shows that services adopt different approaches to user blocking. We have some examples of services immediately blocking for specific kinds of illegal harms, although the evidence does not show how quickly these accounts are identified nor specify how long for. Indeed said that “Between July 1 -December 31, 2021 ... Indeed removed 292,027 accounts due to fraud.”<sup>732</sup> Dropbox’s acceptable use policy covers “material that’s fraudulent, defamatory or misleading or that violates the intellectual property rights of others”, and sharing such content can lead to account actions such as “suspending a user’s access to the Services or terminating an account.”<sup>733</sup>
- 21.24 Broadly, we know it is common for services to ban immediately users who have been found to violate rules around terrorism and child sexual exploitation and abuse. TikTok say 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.”<sup>734</sup> Vimeo’s Acceptable Use Community Guidelines say “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.”<sup>735</sup> X says “in the majority of cases, the consequence for violating our child sexual exploitation policy is immediate and permanent suspension.”<sup>736</sup> It permanently suspends any accounts that violate its violent and hateful entities policy.<sup>737</sup> WhatsApp “ban users when we become aware they are sharing content that exploits or endangers children.”<sup>738</sup> Dropbox says in the case of confirmed CSAM, it disables the user’s account.<sup>739</sup> [CONFIDENTIAL X].<sup>740</sup> Meta says 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.<sup>741</sup>
- 21.25 Ofcom also understands that repeat behaviours can lead to a ban on some services. In theory, when an account reaches the maximum strikes, either generally or in a certain policy area, the account could be banned. **TikTok says, “If an account meets the threshold of strikes within either a product feature or policy it will be permanently banned”.**<sup>742</sup> UK Babe

---

<sup>731</sup>Kim, M., 2022. [85% of Apex Legends Players Responded Better to Direct Feedback Than Outright Bans](#), *IGN*, 23 February [accessed 6 June 2023].

<sup>732</sup> Indeed defines fraud as accounts or job postings that target job seekers or employers with malicious actions or law-breaking activities. Source: Indeed, 2022. [Transparency Report Jul. 1 - Dec. 31, 2021](#), page 7.

<sup>733</sup> Dropbox. [Acceptable Use Policy](#). [accessed 16 August 2023].

<sup>734</sup> TikTok. Supporting creators with an updated account enforcement system, 2023.

<sup>735</sup> Vimeo, 2022. [Vimeo Acceptable Use Community Guidelines](#). [accessed 16 August 2023].

<sup>736</sup> X, 2020. [Child sexual exploitation policy](#). [accessed 24 July 2023].

<sup>737</sup> X, 2023. [Violent and hateful entities policy](#). [accessed 24 July 2023].

<sup>738</sup> WhatsApp. [How WhatsApp Helps Fight Child Exploitation](#). [accessed 24 July 2023].

<sup>739</sup> Dropbox 2022. [Transparency report Jul 2022- Dec 2022](#). [accessed 16 August 2023].

<sup>740</sup> [CONFIDENTIAL X].

<sup>741</sup> Meta response to 2022 Illegal Harms Call for Evidence.

<sup>742</sup> TikTok. Supporting creators with an updated account enforcement system, 2023.

Network says that “accumulating so many points leads to a ban”.<sup>743</sup> [CONFIDENTIAL X].<sup>744</sup> Meta said, “depending on which policy the violating content goes against, your previous history of violations and the number of strikes you have, your account may also be restricted or disabled.” Meta also reported that “for most violations, if you continue to post content that goes against the Facebook Community Standards or Instagram Community Guidelines, despite repeated warnings and restrictions, Meta will disable your account.”<sup>745</sup>

- 21.26 For some harms, a coordinated approach to mass account blocking, rather than manual removal of individual accounts, is used. This often involves a mix of automated detection and human review to identify coordinated campaigns at scale. Meta regularly publishes reports on mass account takedowns for Coordinated Inauthentic Behaviour (which it describes as “when groups of pages or people work together to mislead others about who they are or what they’re doing”).<sup>746 747</sup> As another example, Indeed says, “Depending on the circumstances we work to suspend accounts, most notably with fraudulent actions, and we work through technology and human efforts to identify potentially linked accounts, to remove fraudulent content providers entirely.”<sup>748</sup>

## Effectiveness

- 21.27 We believe that a measure related to strikes and/or blocking has the potential to ensure services comply with their illegal content safety duty in section 10(2) of the Act.
- 21.28 There is evidence to suggest that strike systems can reduce the likelihood of other users encountering illegal content by encouraging offending users to change and stop their illegal behaviour, reducing the likelihood of illegal content being posted on the service. UK Babe Network said that “most people tend to heed” their “warnings” outlined above.<sup>749</sup> Meta shared in an update to its penalty system that it feels people respond to the warning system of strikes, “Our analysis has found that nearly 80% of users with a low number of strikes do not go on to violate our policies again in the next 60 days.”<sup>750</sup>
- 21.29 As set out above, people who post certain types of illegal content do so repeatedly and persistently, implying that blocking may be an effective way to reduce the prevalence of certain types of content on a platform. Similarly, the fact that a wide range of services apply blocking measures in some circumstances implies that such measures are seen as an effective way of addressing online harm.
- 21.30 **However, we note that the effectiveness of strike and blocking systems is largely reliant on how services apply them in practice. This encompasses several aspects including how they vary across different services with different designs, how long services apply them for, and how a service technically enforces them.**

---

<sup>743</sup> UK Babe response to 2022 Illegal Harms Call for Evidence.

<sup>744</sup> [CONFIDENTIAL X].

<sup>745</sup> Meta, 2023. [Taking down violating content](#). [accessed 16 August 2023]; Meta, 2023. [Restricting accounts](#). [accessed 16 August 2023].

<sup>746</sup> Meta, 2022. [Quarterly Adversarial Threat Report](#), page 18.

<sup>747</sup> Meta, 2018. [Coordinated Inauthentic Behavior Explained](#). [accessed 30 May 2023].

<sup>748</sup> Indeed response to 2022 Illegal Harms Call for Evidence.

<sup>749</sup> UK Babe Network response to 2022 Illegal Harms Call for Evidence.

<sup>750</sup> Meta, 2023. [How We’re Improving Facebook’s Penalty System](#). [accessed 30 May 2023].

## Duration of strike or block

21.31 We currently have a limited understanding of how services determine how long strikes and blocking should last. Understanding the duration for which services apply these sanctions, and the reasons that services have chosen those periods, is key to our analysis of what would constitute a proportionate response to different harms. The information we do have indicates there is not a standardised approach across different services. Meta says all strikes on Facebook or Instagram expire after one year.<sup>751</sup> TikTok removes strikes on an account after 90 days, and YouTube similarly says strikes are removed 90 days after issue.<sup>752</sup>

<sup>753</sup>

## How services enforce blocking

21.32 A key aspect of the effectiveness of blocking users is the ability of users operating such accounts to return to a service. For example, Refuge's Marked As Unsafe report suggests that making multiple accounts is a common harassment tactic.<sup>754</sup> We understand that there are a range of approaches services can take to technically enforce a block. For instance, Instagram automatically blocks new user accounts linked to abusive accounts where the same log in details are used.<sup>755</sup> The Online Dating Association says members use "technology related to IP addresses, text and language used, geographical location" to remove users who try and return to a platform after being banned.<sup>756</sup> The Mid-Size Platform Group said many platforms consider ban evasion as being "among the most complex issues platforms must address, and sophisticated bad actors frequently find new ways to evade measures platforms have in place and platforms may deploy a number of tactics to prevent this". It said that members facing this challenge use a range of approaches such as, "detecting patterns of behaviour to prevent users from creating new accounts, investigating reports of ban evasions, mostly using automated detection to effectively identify recidivism."<sup>757</sup>

21.33 One service shared its preference not to ban, to better enable it to monitor user accounts for future illegal activity. A small platform said, "When we discover illegal content and remove it from the site we do not always delete the account associated with that material and monitor the account instead. Sanctions and disincentives will not always lead to less illegal content, and in some cases may make it harder to keep illegal content off of a service."<sup>758</sup> Similarly, Meta said in a 2021 update to its enforcement around CSAM that it was testing new tools and interventions for users who did "not exhibit malicious intent" (for example, sharing content such as viral memes in "outrage or in poor humour").<sup>759</sup> It said it would look to employ a range of interventions, from proactive warnings to removal.<sup>760</sup>

21.34 We are also aware that there are benefits and downsides to different technical methods of enforcing a block. For instance, blocking by username without requiring any other form of

---

<sup>751</sup> Meta, 2022.

<sup>752</sup> TikTok. Supporting creators with an updated account enforcement system, 2023.

<sup>753</sup> YouTube. [Community Guidelines strike basics on YouTube](#). [accessed 6 June 2023].

<sup>754</sup> Refuge, 2022 [Marked As Unsafe](#). Page 9.

<sup>755</sup> Refuge, 2022 [Marked As Unsafe](#). Page 14.

<sup>756</sup> The Online Dating Association response to 2022 Illegal Harms Call for Evidence.

<sup>757</sup> The Mid-Size Platform Group's response to the 2022 Illegal Harms Call for Evidence.

<sup>758</sup> [CONFIDENTIAL X].

<sup>759</sup> Meta, 2021. [Preventing Child Exploitation on Our Apps](#). [accessed 18 April 2023].

<sup>760</sup> Meta, 2021.



identity verification is privacy friendly, but will likely allow the user to return to the platform under a new and potentially similar username. Blocking by email address is also easy to circumvent as users can simply return using a different email address. Blocking based exclusively on IP address is unlikely to be effective given that most internet access services assign “dynamic” (i.e. changing) IP addresses to end-users. This means that blocking a user exclusively on their IP address will only last for as long the user is assigned that IP address, which could be for only a few days or weeks. Further, users who access services using a proxy or VPN are assigned a different IP address each time they connect. We also understand that some mobile network operators use Carrier Grade Network Address Translation (CGNAT) within their network infrastructure which means many mobile internet customers may make use of a single shared IP address. This means that multiple users could find themselves blocked because of the behaviour of one individual.

## Consequences of blocking

- 21.35 We are also conscious that there can be unintentional effects in blocking users. Blocking users from mainstream platforms can have displacement effects, whereby users migrate to alternative platforms with fewer content moderation procedures. While these displacement effects can sometimes result in the toxicity of users’ behaviour increasing, they also tend to reduce the reach of the content. For example, a report on Gab users that created accounts after being blocked on Reddit found that the vast majority became more toxic on their Gab accounts. 40% of those with X accounts either maintained toxicity or got more toxic on Gab. However, the report found that, for the latter group, the Gab accounts had a lower audience reach.<sup>761</sup>
- 21.36 There is a wealth of evidence regarding the effect of blocking networks of ISIS supporters on social media services, and the unintended effects. The evidence suggests that blocking is effective in disrupting supporter networks, although can be associated with radicalising more embedded supporters and fuelling innovation to evade such bans.<sup>762</sup> Disrupting offenders or groups of offenders requires them to rebuild each time, potentially slowing the reach and impact of harmful behaviour, but there is also a known risk of pushing them to other services or spaces. Moonshot, a research group focussed on online extremism and harms, reported that while the threat of blocking can lead some users to moderate their content on mainstream platforms (for example, by using coded language or symbols), it has also led to innovation, and the use of multiple accounts and multiple services to lessen the impact of being blocked.<sup>763</sup>
- 21.37 The risk of migration to other services by those looking to commit terror offences has been highlighted in our Register of Risks, as well as those looking to commit fraud offences.<sup>764</sup>

---

<sup>761</sup> Ali, S., Saeed, M.H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S., Stringhini, G., 2021.

[Understanding the Effect of Deplatforming on Social Networks.](#)

<sup>762</sup> Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson A., Weir, D., 2017. [Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts.](#); Alexander, A., 2017. [DIGITAL DECAY? Tracing Change Over Time Among English-Language Islamic State Sympathizers on Twitter.](#); Dr Elizabeth Pearson argues that among studied ISIS supporters, while suspension can reduce volume of accounts on a service, it can create a sense of community among supporters, who see return from suspension as a shared, bonding experience. Source: Pearson, E., 2018. [Online as the New Frontline: Affect, Gender, and ISIS-Take-Down on Social Media](#), Studies in Conflict & Terrorism, 41 (11), p.19.

<sup>763</sup> From Ofcom engagement with Moonshot, 5 May 2023.

<sup>764</sup> Please see our Register of Risks Volume 2: Chapter 6B (Terrorism Offences), paragraph 74; Chapter 6O (Fraud and financial services offences), paragraphs 36 and 72.



## Rights impacts

### Impacts on users' freedom of expression and freedom of association

- 21.38 As set out in Chapter 12, the right to freedom of expression may be interfered with where to do so is prescribed by law and necessary in a democratic society in pursuit of a legitimate interest. Interferences with the right to freedom of association are subject to a similar test. Given the vast range of harms to which strikes and blocking may apply, any interference with user rights could be in the interests of national security or public safety, 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. However, the potential extent of the impact of strikes and blocking on user rights means that robust evidence and careful analysis is required to be satisfied that any recommendation in this area would be proportionate.
- 21.39 Although blocking and strikes may be a way of tackling illegal content, there are also concerns about the use of these systems on lawful speech. Preventing a user from accessing a service means removing their ability to impart and receive information and to associate with others on that service. It therefore represents, for the duration of the block and in respect of that service, a significant interference with that user's freedom of expression and association. The impact also extends to other users, who will be unable to receive information shared by the blocked user on the service in question. Restricting access to certain functionalities as part of a strikes system may also interfere with user rights, for example if the user is prevented from posting content on the service.
- 21.40 These concerns are more acute if services cannot reliably determine illegal content for the purposes of applying a block or strike. In our 2022 Illegal Harms Call for Evidence, some larger services and civil rights groups noted the difficulty of making determinations of whether individual posts represent illegal content at scale. X said that "It is important to state that determining at scale whether each individual post violates UK law is not only technically infeasible but also, on principle, a decision for the police and courts to make".<sup>765</sup> The Alan Turing Institute suggested, for instance, that "deploying automated measures to tackle illegal content requires the capability to accurately automatically identify this content. When discussing the efficacy of these measures it is important to remember the many challenges automated solutions still face".<sup>766</sup>
- 21.41 We are aware of incidents where strikes and bans have been applied to accounts when content or behaviour has been incorrectly identified as violating platform policies, leading to unintended sanction of users and restriction of their rights:
- a) Instagram users commenting on events in Afghanistan, Israel and Palestine reported having content removed and accounts disabled under the service's "violence and dangerous organisations" policy.<sup>767</sup> Journalists reporting in Tunisia, Syria and Palestine have reported also losing access to Meta under similar policies.<sup>768</sup>

---

<sup>765</sup> X response to 2022 Illegal Harms Call for Evidence.

<sup>766</sup> The Alan Turing Institute response to 2022 Illegal Harms Call for Evidence.

<sup>767</sup> Uddin, R., 2021. [Afghanistan: Muslim Instagram users complain about censorship](#), *Middle East Eye*, 27 August. [accessed 19 May 2023].

<sup>768</sup> Solon, O., 2020. ['Facebook doesn't care': Activists say accounts removed despite Zuckerberg's free-speech stance](#), *NBC*, 15 June. [accessed 6 June 2023].

- b) Multiple sexual health educators reported that TikTok’s ban on nudity and depiction of sexual activities led to their content and accounts being banned, despite platform policies protecting educational content.<sup>769</sup>
- c) In one instance, a parent said they lost access to Google services after pictures taken of their children to demonstrate medical issues for a doctor were flagged and reported to law enforcement as CSAM.<sup>770</sup>

21.42 The Act goes some way towards addressing these concerns, in that it defines illegal content and provides that the threshold for services to take content down is having “reasonable grounds to infer”. More information on this is set out in Ofcom’s proposed Illegal Content Judgements Guidance (ICJG). However, these judgements are, and will remain, inherently difficult in many instances.

### Impacts on users’ privacy

21.43 We currently have limited evidence regarding how services treat users’ personal data for the purposes of strikes and blocking. We would expect services to have in place robust policies to comply with their data protection obligations. Although it is unlikely that a service would need to collect any user data it did not already need to collect for the purposes of offering the service to implement such systems, it is likely that it would be necessary to process and retain some data differently, which increases the risk to users’ privacy.<sup>771</sup>

### Provisional conclusion

21.44 Given the analysis above, we are not proposing to include a broad measure in our Codes which would recommend that services block accounts which post any type of illegal content. There are several reasons for us taking this position:

- a) As the evidence set out above shows, no single system would be suitable for all types of services and harms.
- b) To make a broad recommendation that applies to different types of illegal content, we would need more evidence to ensure that appropriate safeguards are in place to protect user rights. This would require us to set accompanying standards on implementation (including the duration of blocks and strikes), to assess further the effectiveness of technologies used to block users (and for example, any unintended consequences, such as the impact on the freedom of expression of other law-abiding individuals on the same IP address being blocked), and to consider more carefully the impact on users of digital exclusion. We do not yet have sufficient information to develop a position on these points of detail. We are particularly conscious that a broad and loosely defined recommendation could result in over penalisation of users.

---

<sup>769</sup> Iovine, A., 2021. [Why is TikTok removing sex ed videos?](#), *Mashable*, 23 October. [accessed 6 June 2023].  
 Are, C., Briggs, P., 2023. [The Emotional and Financial Impact of De-Platforming on Creators at the Margins](#), *Social Media + Society*, 9(1).

<sup>770</sup> Hill, K., 2022. [A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal](#), *New York Times*, 21 August. [accessed 8 September 2023].

<sup>771</sup> The Council of Europe’s report on content moderation notes the potential data retention around strike policies, “in order to implement measures such as YouTube’s ‘three strikes’ policy, a range of personal and non-personal data must be stored by the company, such as the username of the individual, the name of the complainant, the justification for the removal of the content, dates and times of uploads and removals and so on”. Source: Council of Europe, 2021. [Content Moderation](#), page 30.

- 21.45 We nevertheless consider that the seriousness of users sharing CSAM, the severity of the harm, evidence of the likelihood of repeat offending and multiple indications that many services already impose immediate blocks on users they identify as sharing CSAM means it is warranted to explore a specific blocking measure for users sharing CSAM.
- 21.46 However, we will need to do further work to develop a measure before we can consider proposing its inclusion in our Codes. This further work will need, among other things, to focus on the safeguards needed around the measure and to develop our understanding of the technical challenges associated with blocking.
- 21.47 The consultation questions set out at the start of this chapter invite respondents to provide evidence on the issues set out above. We strongly encourage respondents to engage with these questions where they can so that over the coming months we may consider whether, and if so how, we could recommend a proportionate measure in this space.

## Specific recommendation in respect of proscribed organisations

- 21.48 Proscribed organisations are organisations which have been banned by the UK Home Secretary following assessment against a number of factors set out in legislation, including the specific threat they pose to the UK.<sup>772</sup>
- 21.49 Proscribed organisations differ from all other users in that any activity carried out on an account operated for and on behalf of a proscribed organisation is almost certain to be an offence, and (to the extent it generates content) amount to priority illegal content. Even the setting up of the account would be likely to amount to one or more priority offences. This is because, in addition to the priority terrorism offences relating to proscribed organisations,<sup>773</sup> the priority offence of preparation of terrorist acts captures “any conduct” in preparation for an act of terrorism.<sup>774</sup> An act of terrorism includes any action intentionally taken for the benefit of a proscribed organisation.<sup>775</sup>
- 21.50 As such, we consider that a measure recommending that providers remove accounts operated by or on behalf of a proscribed organisation could be effective in removing this type of illegal content from the service.

## Effectiveness

- 21.51 We recognise that services will rarely be able to be certain that an account is operated by or on behalf of a proscribed organisation. We therefore expect services to remove an account when they have reasonable grounds to infer that it is being run by or on behalf of a

---

<sup>772</sup> Home Office, 2021. [List of proscribed terrorist groups and organisations](#). [accessed 6 June 2023]

<sup>773</sup> 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. Source: Part 2 of the Terrorism Act 2000.

<sup>774</sup> Section 5 of the Terrorism Act 2006; see also section 20(2) for interpretation.

<sup>775</sup> Section 20(2) of the Terrorism Act 2006; and section 1(5) of the Terrorism Act 2000.

proscribed organisation. We consider this to be an appropriate threshold as it is consistent with the threshold in the Act for making an illegal content judgement.

- 21.52 While we believe seeking to discourage users from returning to a service is an important consideration for the effectiveness of a user access measure, methods to prevent return is an area where we are still developing evidence. However, any disruption of the activities of proscribed organisations is likely to be beneficial. Removing the account disrupts their activities by reducing their ability to communicate with their followers, at least for a period. If a user did return to a platform, the removing of their original account would still require them to spend time rebuilding networks with other users, adding burden to the process. We are therefore at this stage only recommending that services remove the account of the user in question.
- 21.53 Although we consider this measure would in principle be effective at tackling the harm caused by proscribed organisations, we recognise that identifying an account as being operated by or on behalf of a proscribed organisation presents challenges.
- 21.54 When a service concludes that a piece of content either amounts to a proscribed organisation offence, or is in breach of an equivalent standard as set out in their terms and conditions, or when it has received a complaint or been made aware of an account that an account may be operated by or on behalf of a proscribed group, we expect the service will then proceed to consider whether the account might be operated by or on behalf of that proscribed organisation. In many cases, services are likely to become aware of an account linked to one of these groups due to a piece of flagged illegal content through its moderation process, although in some cases it may be law enforcement, another user or a member of public who identifies the content and requests it to be taken down via reporting or complaints processes. In some cases, proactive detection of potentially illegal behaviour may be used, for example if the service considers it necessary to prevent reoffending, but this is not a requirement of this mitigation.
- 21.55 When considering content, services may refer to the list linked in our ICJG or the government's own website to find proscribed organisations.
- 21.56 When looking at an account, we expect services to remove the account where they have reasonable grounds to infer that it is operated by or on behalf of a proscribed group. There are several factors that we provisionally consider may give rise to reasonable grounds to infer. These include where a combination of the following user profile factors is present:
- a) **Username:** The username may be, contain, or make reference to that of a proscribed organisation or a known/listed alias for a proscribed organisation.<sup>776</sup>
  - b) **User profile images such as profile/account/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.
  - c) **User profile information:** Other information fields attached to the account could suggest the account is operated by or on behalf of a proscribed organisation. This may include the name of the organisation included in a user 'bio', or another descriptive field such as those describing education, workplace or political beliefs.

---

<sup>776</sup> Schedule 2 to the Terrorism Act 2000 and the official Government list of proscribed organisations contain aliases of proscribed groups recognised by the Secretary of State. Source: Home Office, 2021.

- 21.57 We recognise that the above factors may not be present, but an account may nevertheless be operated by or on behalf of a proscribed group. As such, we consider that reasonable grounds to infer may also arise where a significant proportion of a reasonably sized sample of the content recently posted by the user amounts to a proscribed organisation offence. We do not consider it practicable to specify precisely how much content a service should consider for this purpose, as we expect that this would vary both from service to service and from case to case.
- 21.58 For the purposes of the preceding paragraph, “content” does not include content that has been privately communicated, unless the relevant service has explicit consent to view the content in question, for example having received a report about a private communication. This is because of the implications for a user’s right to privacy of a service viewing privately communicated content without explicit consent. We discuss in more detail the privacy implications of this measure below.

## Costs and risks

- 21.59 The costs that this measure can potentially involve include: designing a process for staff to follow; providing associated training to staff on this process; staff then assessing any accounts under suspicion; staff removing an account as necessary; and any costs if a user were to appeal a decision to take down an account. The costs will depend on how the service decides to approach implementing this measure and the methods it uses to remove users from the service.
- 21.60 However, these costs will not be relevant for all services. On most services, and especially small services, we consider it unlikely that there are accounts operated by proscribed organisations, nor do we expect there to be any illegal terrorist content. In these cases, we do not envisage such services needing to incur any costs in advance of receiving a complaint or otherwise becoming aware that a proscribed group may have an account on their service. Such services would not need to train staff in advance or develop processes, and would only incur the costs of assessing an account in the unlikely event they were made aware of one. And if they never receive any such complaint or otherwise become aware of a suspicious account, they would not incur any one-off or ongoing costs related to this measure.<sup>777</sup>
- 21.61 In contrast, for other services, content by proscribed organisations is more likely. For example, larger social media services may need to incur more greater costs. The costs will depend on how the service decides to approach implementing this measure and methods it uses to remove users from the service. Some such services may consider it appropriate to set out a process for its staff on how to assess whether an account is operated by or on behalf of a proscribed organisation (including considering factors set out in paragraph 21.56). This is likely to require regulatory and/or legal staff input and costs of this are likely to vary depending on size and type of service. It will then be necessary to train staff to recognise a proscribed organisation and provide them with appropriate materials, to confirm that an account should be removed from the service. However, as section 10(3)(b)

---

<sup>777</sup> To maximise reach and impact we expect these groups to focus on a few accounts and specific services as accounts need to be easily identifiable to be effective in recruiting and/or spreading the message for the cause.<sup>778</sup> The additional step could add a couple of hours to one-off training time adding less than £200 per trained employee to the cost of attending training. In addition, there will be costs relating to providing this training, and the service may choose to provide some top-up training for its content moderators on an ongoing basis.

of the Act requires services to swiftly take down any illegal content as they become aware of it, and services are likely to train their staff to do this, adding an additional step to identify a proscribed group is unlikely to incur significant additional costs.<sup>778</sup>

- 21.62 The account reviews and take downs will require content moderators to review and assess whether an account is operated by a proscribed organisation and potentially other second level support staff to remove the account. We do not expect account removals to be particularly costly or complex or require technical expertise in most cases. We recognise that services may use various ways to remove accounts. Alternatively, a service may consider it appropriate to automate an account removal processes, which is likely to incur higher upfront costs while ongoing costs may be lower as a result.
- 21.63 When staff need to review and remove an account, we assume this takes two hours of a content moderator's time and one hour of a software engineer's time; based on these assumptions this would impact a per account reviewed and removal of ~£140.<sup>779</sup> We think this estimate is on a high side and most likely exaggerates the cost of removing an account, and in practice expect this cost to be lower in most cases.<sup>780</sup> The ongoing costs of this measure would depend on how commonly terrorist content is shared on a service, as detecting this content, along with user complaints, may prompt a review of an account.
- 21.64 On those larger services illegal terrorist content may be more common, more accounts will be flagged and removed, although the prevalence by service is likely to vary.<sup>781</sup> While the total cost of manually removing users would be higher in these cases, the benefits are also bigger in terms of a reduction in the spread of illegal content and recruitment of people to the proscribed organisation. Furthermore, many of these services already block users for sharing illegal terror content and already have systems and processes in place which would suggest that the incremental costs of this mitigation may be limited (although the service may have to consider whether its existing processes are compliant with our proposal). In addition, it is possible that there are some common costs with our other proposals (e.g.

---

<sup>778</sup> The additional step could add a couple of hours to one-off training time adding less than £200 per trained employee to the cost of attending training. In addition, there will be costs relating to providing this training, and the service may choose to provide some top-up training for its content moderators on an ongoing basis.

<sup>779</sup> Based on the labour cost assumptions set out in Annex 14.

<sup>780</sup> However, 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.

<sup>781</sup> For example, Snap Inc. removed 73 and 132 accounts for violating their policy prohibiting terrorist and violent extremist content in six months to 30 June 2022 and 31 December 2022 respectively. These numbers reflect Snap Inc's definitions and are based on global figures, which may not reflect the UK situation (although it is possible that all terrorist content violations on a service prompt a suspicion of an account being ran for by or on behalf of proscribed groups and require an investigation by the service, even if these are not identified as such by the Home Office). Furthermore, not all these accounts are run by or on behalf of proscribed terrorist organisations, which means that the number of accounts taken down may be lower. Based on Snap Inc's blocked accounts in 2022, and an estimated cost of ~£140 per account take down, removing users because of this mitigation could cost ~£29,000. In contrast, X suspended ~45,000 and 34,000 unique accounts for violating its prohibition policy of promoting terrorism and violent extremism in the six months to June and December 2021 respectively. Based on the prevalence of terrorist content on X in 2021, and if all these accounts were by proscribed groups, using our estimate of taking down an account of ~£140, assessing and removing these users would cost ~£11m. Sources: Snap Inc., 2022. Transparency Report [January 1, 2022 – June 30, 2022](#). [accessed 16 August 2023]; Snap Inc., 2023. Transparency Report [July 1, 2022 –December 31, 2022](#). [accessed 8 September 2023]; X, 2022. [Rules enforcement January - June 2021](#). [accessed 16 August 2023]; X, 2022. [Rules Enforcement July -December 2021](#). [accessed 16 August 2023].



content moderation) if implemented together, creating some cost efficiencies for the service.<sup>782</sup>

- 21.65 Finally, any service that removes an account may need to incur costs if a user were to appeal a decision to take down an account. While establishing such an appeal process is a requirement in the Act, processing complaints relating to this measure where the service has wrongly removed a user would incur incremental costs to the service. This would include a cost linked to reviewing the appeal and restoring the account if the appeal is legitimate (which is likely to be similar to the cost of takedown discussed earlier, although the review time by content moderator or other staff per case could be longer). We expect the volume of such appeals to be low, as the number of accounts blocked in most cases is likely to be low, although information on appeal numbers is limited.<sup>783</sup> Overall, we do not expect appeals that occur as a result of this mitigation to impose substantial additional costs on services.

## Rights impacts

### Freedom of expression and freedom of association

- 21.66 As set out previously, any interference with the right to freedom of expression or freedom of association must be prescribed by law and necessary for a legitimate aim. In the case of proscribed organisations, the interests of national security or public safety and the prevention of crime are most relevant. In order to be ‘necessary’, the restriction must correspond to a pressing social need, and it must be proportionate to the legitimate aim pursued.
- 21.67 Removing 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. Doing so could reduce illegal activities by such groups, like recruitment and seeking of support for such groups.
- 21.68 Given that a terrorist organisation’s purposes are fundamentally inconsistent with democracy and human rights, and that they do not usually have legal personality, it is not clear that an organisation which has been proscribed would necessarily have rights to free expression or free association that can be interfered with. If it does, the impact of proscription on those rights is already taken into account in the decision of the Home Secretary to proscribe it. Therefore, the concerns set out above as to the impact of blocking and strikes on human rights do not arise for correctly identified proscribed organisations as they do for other users. As a proscribed organisation should be removed for as long as it is proscribed, there are also not the same difficulties in determining how long it is proportionate for a block to last.

---

<sup>782</sup> For example, there are likely to be some similarities in the required system changes services may have to take to implement our proposed blocking and content moderation mitigations, which in turn could mean some overlap in upfront implementation and ongoing costs.

<sup>783</sup> For example, in Q4 2022, Discord reported that ~15% of account holders who were disabled for violent extremism appealed (fewer than 0.2% of those were successful). Appeals as a proportion of accounts removed fell gradually in 2022 from ~28% in Q1 2022. Discord, 2023. [Transparency Report 2022 Q4](#). [accessed 16 August 2023]; Discord, 2022. [Transparency Report 2022 Q3](#). [accessed 16 August 2023]; Discord, 2022. [Transparency Report 2022 Q2](#) [accessed 16 August 2023]; Discord, 2022. [Transparency Report 2022 Q1](#). [accessed 16 August 2023].



- 21.69 That said, we recognise that there is a risk to users' human rights if their accounts are incorrectly identified as being operated for or on behalf of a proscribed organisation and consequently blocked. However, we are consulting on what we see as a cautious approach to identifying such accounts in practice.
- 21.70 A further means of safeguarding against the risk to human rights of incorrect identification is that the Act requires services to take appropriate action in response to certain complaints, including appeals by UK users (see section 21(4)(d)). This obligation only applies if a service has made an illegal content judgement. We recognise that services' terms and conditions may consider a wider range of content to be terrorist content than as defined in domestic law. However, given the broad nature of the offence of preparing a terrorist act (as discussed at paragraph 21.49 above), we consider that any content takedown decision (including removing an account) based on content being related to a proscribed organisation is likely to be an illegal content judgement within the meaning of the Act. The complaints obligation would therefore apply, enabling UK users who believe they have been wrongfully blocked to appeal.
- 21.71 Users' rights would be affected during the period in between being removed and their appeal being considered. However, we consider this to be proportionate in the circumstances. As above, a service would have needed to have established a combination of factors before the account was removed.

## Privacy

- 21.72 We recognise that the implementation of this measure, in particular a service reviewing content posted by a user, could also have implications for users' right to privacy. An interference with privacy must be in accordance with the law and necessary for a legitimate aim.
- 21.73 Similarly to freedom of expression, it is not clear that proscribed organisations have any right to privacy because they often do not have legal personality. We acknowledge, however, that the actions which could amount to an infringement of privacy would, by definition, be taking place in relation to a user who had not yet been – and may never be – identified as running the account for or on behalf of a proscribed organisation. It has to be assumed that the user concerned does, therefore, have a right to privacy which could be infringed.
- 21.74 The recommendation we are considering is that services should consider whether the account is run for or on behalf of a proscribed organisation when they have identified content posted to the account that either amounts to a proscribed organisation offence, or is in breach of an equivalent standard as set out in their terms and conditions, or when it has received a complaint or been made aware of an account that an account may be operated by or on behalf of a proscribed group. It would therefore only be triggered when there were grounds to suspect that the account may be run for and on behalf of a proscribed organisation. Given the level of risk such organisations pose to public safety and national security, as well as the fact that running such an account would by definition be a crime, we consider that the interference is likely to be proportionate in cases where services were considering content communicated publicly. While we consider it possible for a person to have a reasonable expectation of privacy in such content, it is less likely that they will.
- 21.75 We do not consider it practicable to specify precisely how much content a service should consider for this purpose, as we expect that this will vary both from service to service and

from case to case. However, we consider that recommending that services take only a ‘reasonably sized sample’ of ‘recent content’ makes it clear that it is not necessary for services to review all the content posted by the account concerned in order to follow this measure. This should help to ensure that any interference with an affected user’s privacy is proportionate. More generally, service will remain subject to general privacy and data protection laws in determining how much to review.

- 21.76 Our provisional position is that it would not be proportionate to expect services to review content communicated privately unless they have explicit consent to review particular content.
- 21.77 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 users’ right to privacy would likely be proportionate.

## Provisional conclusion

- 21.78 We propose to recommend for our Code on terrorism the following measure requiring services to remove accounts operated by or on behalf of a proscribed terrorist group:
- a) Services should remove a user account from the service where they have reasonable grounds to infer it is operated by or on behalf of a proscribed organisation.
- 21.79 We consider that this measure would deliver significant benefits. Content posted by proscribed organisations can result in significant harm to people in the UK and beyond. Stopping proscribed organisations using accounts on user to user services has the potential to disrupt their activities and reduce their ability to disseminate terrorist content.
- 21.80 The ongoing costs of this measure are likely to depend on how commonly terrorist content is shared on a service, as detecting this content, along with user complaints, may prompt a review of an account. As a result, the costs of this measure are likely to scale with the benefits. While terrorist organisations can sometimes target small services because they have fewer resources to moderate content,<sup>784</sup> this is only likely to happen to a small fraction of the total number of small services. We anticipate that the large majority of smaller services will not need to incur any costs as a result of this measure. But where a small service is targeted, it would need to incur the costs of reviewing and removing accounts. Given the severe nature of the harm caused by proscribed organisations, we consider that the benefits flowing from this will be large enough to justify the costs of the measure. We therefore consider that there is a strong case for applying this measure to all services regardless of size.
- 21.81 We recognise that, when implementing this measure, services will most likely review content posted by accounts operated by users, including users who are not 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 above, and in particular the potential for this measure to mitigate the significant risk of harm posed by accounts run by proscribed organisations, we consider the interference would be proportionate.
- 21.82 We are not making specific recommendations for how services should implement our proposed measure and consider that services should do this in a way that is most

---

<sup>784</sup> Please see our Register of Risks Volume 2: Chapter 6B Terrorism Offences, paragraphs 73 to 76.

appropriate for them. We think there is sufficient flexibility to allow all services to approach this mitigation in a proportionate way.

## Verifying users' identity

---

### Harms that this measure seeks to address

- 21.83 As set out in our Register of Risks, the ability to make user profiles an “anonymous user profile”, and providing a feeling of pseudonymity or ‘anonymity’, can potentially embolden users to engage in harmful behaviours, including threats harassment and stalking, hate speech, CSEA, foreign interference, extreme pornography, intimate image abuse.<sup>785 786</sup> In regard to CSEA, the Register of Risks notes that anonymity or pseudonymity can hinder legal investigations, or enable users to pretend they are someone they are not to, for example, gain access to children for the purposes of sexual grooming.<sup>787</sup> It also highlights how the ability to create multiple accounts can facilitate illegal harms such as intimate image abuse.<sup>788</sup> We also note how the ability to upload pseudonymously or ‘anonymously’ can facilitate said harms.
- 21.84 There is evidence to suggest that a user feeling that their identity is somewhat<sup>789</sup> or entirely hidden can facilitate illegal harms through a ‘disinhibiting effect’, which causes users to “act out”.<sup>790</sup> This effect has been described as a risk factor by civil society groups working to mitigate online harms.<sup>791</sup> Civil society groups highlight the importance of services ‘knowing your client’ to prevent such harms.
- 21.85 The ability to make accounts without requiring the identity or a high level of assurance of the identity of an account holder can also lead to harms facilitated by users impersonating another user or holding multiple accounts. We examine the effect of this for notable users in Chapter 20 on Enhanced User Controls.

---

<sup>785</sup> “Anonymity” refers to when a person cannot be identified or singled out from any other individual. The term is often conflated with “pseudonymity”, where persons are distinguished from one another with aliases that are not linked to their real-world identities. Obtaining full anonymity, where an account can never directly be linked to an individual’s real-world identity, is almost never possible on an online platform (which can use identifying features such as IP or email addresses). For users, interacting with pseudonymous or anonymous users can feel the same; users may also feel anonymous even when they are identifiable (or potentially identifiable) by a service.

<sup>786</sup> Please see our Register of Risks Volume 2: Chapter 6E Harassment, stalking, threats and abuse offences, paragraphs 56 to 59; Chapter 6F Hate offences, paragraphs 43 to 46; Chapter 6C Child Sexual Exploitation and Abuse, section Child Sexual Abuse Material, paragraph 51; Chapter 6P Foreign Interference Offence, paragraphs 52 and 53; Chapter 6L Extreme Pornography offence, paragraph 31; Chapter 6M Intimate Image Abuse offences, paragraphs 51 and 52.

<sup>787</sup> Please see our Register of Risks chapter Child Sexual Exploitation and Abuse, section Grooming, paragraphs 44 to 46.

<sup>788</sup> Please see Our Register Of Risks Volume 2: Chapter 6M Intimate Image Abuse, paragraph 50.

<sup>789</sup> An identity may be somewhat hidden where, for example, it is not visible to other users but is known by the provider of the service.

<sup>790</sup> Suler, J., 2004. [The Online Disinhibition Effect](#), *Cyberpsychology & behavior*, 7 (3).

<sup>791</sup> Clean Up the Internet (Babbs, D.), 2020. [Time to take off their masks? How tackling the misuse of anonymity on social media would improve online discourse and reduce abuse and misinformation](#); Antisemitism Policy Trust, 2020. [Regulating Online Harms: TACKLING ANONYMOUS HATE](#).

## Options

21.86 Below we consider whether there is a case for recommending that services deploy identity verification ('IDV') as a potential mitigation against the risks posed by users feeling anonymous online. IDV is the process of a service confirming that a user is the person they claim to be or possesses an attribute they claim to have.<sup>792</sup>

### Discussion of current use and the efficacy of identity verification in deterring illegal content

21.87 We have considered whether requiring users to provide details that enable services to establish their real-world identity or government identity could effectively deter users from posting illegal content or behaving in illegal ways on the service. Providing identifying details could limit illegal content in ways including:

- a) Users may be deterred from sharing illegal content if they are aware that they are connected to their real world identity, and that this could be shared with law enforcement; and
- b) A high-level of assurance in identity verification may also make it harder for users to circumvent platform enforcement.

21.88 As such, measures requiring services to establish a user's identity could potentially assist services in complying with the illegal content safety duty in section 10(2) of the Act. However, as we go on to explore in more detail below, based on the evidence we consulted we do not believe that the benefits of recommending a Code measure requiring services to adopt IDV to tackle illegal harms would justify the potential impacts on users' privacy and freedom of expression.

## Effectiveness

21.89 We are aware that many services use IDV for different levels of user access controls and in different use cases. In terms of how effective these measures are in preventing harm from illegal content, we have some evidence suggesting that IDV can reduce the level of illegal content available on services hosting pornographic content. We have engaged with a U2U adult service provider that has implemented real-world IDV measures at point of upload, which believes its introduction of verification, alongside other mitigations, contributed to a reduction in harm and illegal content. In engagement with Aylo (formerly know as MindGeek) in November 2022, Ofcom were told "No changes noted in visitor numbers due to the evolution and addition of trust & safety technology, but ID measures since January 2021 are deterring illegal uploads."<sup>793</sup> In a meeting with Aylo in October 2021, it said that

---

<sup>792</sup> There are various methods of identity verification, and solutions are continuously developing. The identity verification process usually starts with gathering from the user certain claimed attributes and evidence of these, followed by validating that evidence, then establishing that the person with those attributes is the one seeking access. Some methods offer lower levels of assurance, such as those that only require email addresses and/or telephone numbers. Some offer higher levels of assurance, such as those that require government-issued passports or other documentation that links to someone's real world identity. Higher assurance may present a trade off in other areas, including user privacy or inclusivity. There is no method that is 100% accurate. Source: National Institute of Standards and Technology, 2017. [NIST Special Publication 800-63A, Digital Identity Guidelines, Enrollment and Identity Proofing](#). [accessed 16 August 2023]; Cabinet Office and Government Digital Service, 2023. [How to prove and verify someone's identity](#). [accessed 5 June 2023]

<sup>793</sup> Meeting with Aylo (Mindgeek), 2<sup>nd</sup> November 2022.

“When comparing the first 6 months after the changes to the last 6 months prior, and factoring in the reduction in uploads, Aylo have seen a relative 55% decrease in attempted violative content uploads since it introduced the uploader ID requirements.”<sup>794</sup>

- 21.90 The National Center for Missing and Exploited Children’s CyberTipline yearly report showed a drop in CSAM content reported by adult service provider MindGeek from 2020 to 2021, which coincided with the introduction of tools to tackle CSAM on service, including IDV at upload.<sup>795 796</sup>
- 21.91 However, our evidence on the efficacy of IDV measures is not conclusive:
- a) We know people carry out online illegal behaviour while identifiable to other users and the service. X’s report of abuse towards football players following the 2020 Euros said 99% of accounts involved could be identified.<sup>797</sup> A report by Revealing Reality for the Department of Culture, Media and Sport provided an overview of research and global identification pilots, and concluded, “Even where it looks as though there is a link, isolating the role that anonymity plays in facilitating or magnifying abuse is practically impossible...removing anonymity is rarely suggested as the best solution to reducing abuse” and suggested looking to other mitigations, such as “limiting the disposability of accounts.”<sup>798</sup>
  - b) Often identity verification is used in tandem with other functionalities to prevent the upload of illegal material. This means it is difficult to disentangle the effect of verification from other measures implemented. For example, MindGeek’s transparency reports have outlined a range of technologies and policies introduced to deter harmful content alongside identity verification for uploaders, including deterrence messaging, hash matching and trusted flaggers.<sup>799</sup> Make Love Not Porn have a team of curators that review submissions.<sup>800</sup> MindGeek’s 2022 transparency report indicates it scans all photo and video content using a range of technologies, including against hash databases.<sup>801</sup>
  - c) While ID verification requirements can theoretically deter or slow some users looking to upload illegal content, their efficacy is likely to rely on the robustness of the process involved in verifying users. There is at least one reported account of users managing to be approved for access via identity verification processes while using forged or not their own documents.<sup>802</sup>

## Rights impacts

- 21.92 Measures taken by services to require identity verification are likely to affect users’ right to respect for their private life and correspondence. We recognise that different identity verification methods offer different levels of assurance, and consequentially may result in a

---

<sup>794</sup> Meeting with Aylo 6<sup>th</sup> October 2021

<sup>795</sup> National Centre For Missing and Exploited Children, 2020. [2020 CyberTipline Reports by Electronic Service Providers](#), page 4.

<sup>796</sup> National Centre For Missing and Exploited Children, 2021. [2021 CyberTipline Reports by Electronic Service Providers](#), page 4.

<sup>797</sup> X, 2021. [Combatting online racist abuse: an update following the Euros](#). [accessed 16 August 2023].

<sup>798</sup> Revealing Reality, 2022. [Abuse and anonymity](#).

<sup>799</sup> MindGeek, 2021. [Transparency report](#).

<sup>800</sup> Make Love Not Porn. [Frequently Asked Questions](#). [accessed 20 July 2023].

<sup>801</sup> MindGeek, 2022. [Transparency report](#).

<sup>802</sup> Titheradge, N. and Croxford, R., 2021. [OnlyFans must do more to protect children, watchdog says](#). *BBC News*, 11 June. [accessed 16 August 2023].

differing level of privacy intrusion. The usual trade-off is that the higher the level of assurance, the more intrusive the method is likely to be and the higher its impact on privacy.

- 21.93 Irrespective of the approach taken, the impact on user privacy may be mitigated by services having in place robust measures to comply with their data protection obligations, including requirements to keep data by which users can be identified for no longer than is necessary for the purposes for which the data was processed. However, given the intended deterrent effect of an identity verification measure as discussed above, services could end up retaining such data indefinitely, presenting a greater intrusion into users' right to privacy.
- 21.94 Anonymity online is also an important enabler of freedom of expression. For example, being able to speak anonymously enables individuals to express themselves without fear of repercussion from employers, insurers, family members or their community. Being able to receive such communications enables everyone to become aware of information and ideas which may otherwise not become public. This is particularly true of political and journalistic speech, which are afforded the greatest degree of protection by the law.
- 21.95 Certain groups particularly benefit from anonymity, for example:
- a) Whistle-blowers, journalists, and activists are often highlighted as those benefitting from anonymity to carry out their work. These individuals often benefit from anonymity to avoid persecution for speaking politically and journalistically, which, as noted above, are the most protected forms of speech. In turn, everyone in society can receive the information and ideas they impart.
  - b) The ability to exist online without identification also benefits some minority groups that use platforms to connect and share experiences, but also to mobilise and exercise their right to freedom of association.
  - c) Victims and survivors of illegal harms, particularly sexual harms and domestic abuse, can also benefit from spaces that do not require identification. Our Register of Risks highlights the need for victims of gendered abuse to feel anonymous and some victims have shared their efforts to conceal their identity online.<sup>803</sup> Support groups have encouraged victims to protect their real-world identity online. A level of anonymity between users can also be extremely important for victims and survivors of child sexual abuse who want to seek the support or advice of other survivors in online forums or networks, without disclosing their identity.
- 21.96 Requiring users to verify their identity to access or speak on a service could therefore infringe their right to privacy, freedom of association and freedom of expression. However, these rights are not absolute rights, such that an interference with them may be lawful if it is in accordance with the law (privacy) or prescribed by law (freedom of expression and freedom of association) and necessary in a democratic society for the pursuit of a legitimate interest, if the interference is proportionate to the aim pursued.
- 21.97 Given the broad range of illegal harms of which anonymity increases the risk, any interference could be said to be in pursuit of several aims, including the prevention of disorder or crime, the protection of the rights and freedoms of others, and the interests of national security. We recognise that for these reasons some services may choose to introduce IDV of their own volition.

---

<sup>803</sup> Please see our Register of Risks , Volume 2: Chapter 6E Harassment, stalking, threats and abuse offences, paragraph 56.



## Provisional conclusion

- 21.98 Taking together the considerations above, we do not propose to recommend in our Codes that services implement identity verification to mitigate illegal harms.
- 21.99 We do have limited evidence to suggest that identity verification can act as a deterrent for users looking to engage in illegal behaviour, and several services have implemented user verification with this in mind.
- 21.100 However, the evidence of the efficacy of user verification in deterring illegal content is mixed. In addition to there being important benefits to anonymity for some groups, there are also user rights implications associated with identity verification.
- 21.101 As such, we consider that we are currently unable to assess the proportionality of a recommendation that services apply any sort of IDV measure to comply with the illegal content safety duty in section 10(2) of the Act. Given our provisional conclusion not to recommend any IDV measure for the reasons set out above, we have not considered the potential costs of such a measure. If we return to this area in the future to make recommendations, we will at that point consider costs.
- 21.102 We note, however, that we are proposing other measures in the Codes which may similarly act as a deterrent to illegal behaviour.
- 21.103 We also note that the Act imposes on services a specific child protection duty to use proportionate systems and processes designed to prevent children of any age from encountering content specified in regulations by the Secretary of State (section 12(3) of the Act). Services must comply with this duty by adopting highly effective means of verifying or estimating the age of users. We expect to return to this issue as part of our Protection of Children consultation.
- 21.104 For services that publish or display pornographic content on its service, i.e. content in scope of Part 5 of the Act, specific requirements preventing children from not normally being able to encounter this content are in place. Pornographic services that are U2U or search have a different set of requirements, which will be set out in Codes and will be consulted on as part of our protection of children consultation.
- 21.105 As stated at the beginning of this chapter, verification for user empowerment will also be explored in later phases of our work.

## Verifying users' age

---

- 21.106 Age assurance is an umbrella term to capture the range of measures that can be used to establish a user's age, including age verification and age estimation. There are a range of tools available that can verify a user's age. The key use-case of age assurance is to ensure services understand which users are children and which users are adults on their service, with the aim of protecting children from harm online while allowing adults to access legal content.
- 21.107 As noted above, under the Act services must use age verification or estimation to comply with the duty to use proportionate systems and processes designed to prevent children of any age from encountering content specified in regulations by the Secretary of State. Further, requiring users to verify their age has the potential to prevent children from being exposed to other illegal harms, including grooming (as discussed in our Register of Risks). However, it may also inadvertently block children's access to age-appropriate online spaces.



- 21.108 There are a range of age assurance techniques available which are capable of achieving varying degrees of accuracy and effectiveness. In addition to understanding what age assurance technology is capable of technically achieving, the potential for these technologies to impact on user rights must be considered. We are continuing to build our evidence base in this area in relation to available technologies and expect to return to this matter in our work focusing on protecting children online. In relation to provider pornographic content, Ofcom has a duty to provide guidance to inform services' compliance decisions. We will be consulting on this Guidance shortly.
- 21.109 We are aware that some services already voluntarily collect information about users' age, for example to inform targeted advertising. In Chapter 18, we consider this evidence in the context of our proposed recommendations on default settings for child accounts and support for child users.
- 21.110 We will be publishing two consultations in the coming months that will address our approach on age assurance; our part five guidance later this year, and our Protection of Children consultation next year.

## 22. Search features, functionalities and user support

### What is this chapter about?

This chapter sets out our proposals for measures search services can take to design their services in such a way as to protect people from harm.

### What are we proposing?

We are making the following proposals for all large general search services:

- **Services that use a predictive search functionality should offer users with a means to easily report predictive search suggestions which they believe can direct users towards priority illegal content.** When a report is received, services should consider whether the wording of a reported predictive search suggestion presents a clear and logical risk of users encountering search content that is priority illegal content. If a risk is identified, services should take appropriate steps to ensure that the reported predictive search suggestion is not recommended to any user.
- **Services 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.** This information should include a helpline and links to freely available supportive information provided by a reputable mental health or suicide prevention organisation. It should also be prominently displayed to users in the search results.
- **Services should employ means to detect and provide warnings in response to search requests the wording of which clearly suggests that the user may be seeking to encounter CSAM.** This warning should include information about the illegality of CSAM and links to resources provided by a reputable child sexual abuse organisation to help users refrain from committing CSEA offences. It should also be prominently displayed to users in the search results.

### Why are we proposing this?

Predictive search functions can sometimes suggest search terms which lead users to harmful and potentially illegal content. The first measure we have proposed would help address this problem. The evidence we have assessed suggests that the second two measures could reduce the probability of users encountering suicide promotion content and CSAM respectively.

The measures we are proposing largely reflect what we understand to be current industry standard practice. We note that the publicly available evidence base on search services is relatively limited. Therefore, at this stage, we are focusing on codifying a small number of elements of established best practice rather than pushing for material changes in search services' safety procedures. As we learn over time, we expect to build on and refine our approach.

### What input do we want from stakeholders?

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

## Introduction

---

- 22.1 In this chapter, we focus on search features, functionalities and user support measures. The Act sets out that, in meeting their safety duties in clause 23, search services should take steps, where proportionate, relevant to:
- a) the “design of functionalities, algorithms and other features relating to the search engine” (section 27(4)(b));
  - b) “functionalities allowing users to control the content they encounter in search results” (section 27(4)(c));
  - c) “content prioritisation” (section 27(4)(d)); and
  - d) “user support measures” (section 27(4)(e)).
- 22.2 Search services are distinct from U2U services in that they do not facilitate the sharing or uploading of content by the user of the service but rather facilitate access to more than one website or database. As such, search services can act as a gateway to illegal content that is present elsewhere online. There is evidence that general search services can be used to access content related to a wide range of offences, including, amongst other things, terrorism, hate, extreme pornography, CSAM (Child Sexual Abuse Material) and fraud.<sup>804</sup> However, there is a scarcity of verifiable evidence on the efficacy of existing safety measures that are used by search services to protect their users from harm. In light of this scarcity of evidence, the recommendations we consider in this chapter are largely reflective of current industry practice.
- 22.3 As set out in Chapter 11, we distinguish between the following types of search services: general search services (which enable users to search the web by inputting search requests on any topic) and vertical search services (which focus only on a specific topic or genre of content). Within general search we also distinguish between services that only rely on their own indexing and those which contract to obtain search results (which we call downstream general search services). A longer description of each of these types of service can be found in paragraph 11.65.
- 22.4 Downstream general search services, which buy the index from another general search service, may not have direct control over the results that they display to users and therefore may not be able to directly implement a measure that would require changes to how search results are indexed. The level of control that a downstream general search service has over the index depends on the contract the provider has with the service they buy the index from. The details of these contracts is not publicly known and is likely to differ from service to service.<sup>805</sup>

---

<sup>804</sup> Volume 2: Chapter 6C CSEA (grooming and CSAM), paragraphs 6C.22-6C.27

<sup>805</sup> We are aware that, in its advertising market study, the CMA said none of the contracts it had looked at allowed the downstream general search service to re-rank the search results they received from Google or Bing. Source: CMA, 2020. [Online platforms and digital advertising: Market study final report](#), Box 3.3 page 97 and paragraph 3.85 [accessed 29 September 2023].

- 22.5 It may therefore be necessary for downstream general search services to ensure that the index they buy from other providers in order to provide their search service reflects the most up to date changes those providers have made to comply with the safety duties themselves.
- 22.6 There is no clear evidence to suggest that vertical search services play a significant role in the dissemination of priority illegal content or other illegal content. We have assessed vertical search services as having a low risk profile and we are therefore minded to exclude these services from the scope of the measures proposed in this chapter.<sup>806</sup>
- 22.7 After describing the three measures we propose, we describe another measure we considered and explain why we are not proposing it at this stage.

## Measure relating to the design of predictive search functionalities in search services

---

### Harms this measure seeks to address

- 22.8 Predictive search functionalities are used by several search services in their search engines, particularly large general search services, such as Google’s autocomplete functionality and Microsoft Bing’s autosuggest tool. Predictive search functionalities are algorithmic features that are embedded in the search bar. When a user begins to input a search request, the algorithm predicts the search and suggests possible related search terms. Predictions are based on many factors including past and other user queries, location and trends.
- 22.9 While predictive search can be a helpful and time-saving tool for users, search prediction has been identified as a risk factor for search services in Ofcom’s Register of Risks. For an explanation of search prediction as a risk factor, please see Chapter 6.3, paragraphs 6.37 - 6.39.
- 22.10 We see predictive search as a general search service functionality that can increase the risk of individuals encountering search content that is illegal content. Specifically, there is a risk that a search prediction may lead a user to illegal content that they might otherwise not have encountered had the search suggestion not been surfaced.
- 22.11 Ofcom research published in September 2023 on the accessibility via search services of articles and items for use in the commission of fraud, found that autocomplete suggestions and ‘related searches’ can help users find prohibited fraud-related content by recommending more detailed or accurate search suggestions for the kind of prohibited articles or items that a user might be searching for.<sup>807</sup>
- 22.12 Respondents to our 2022 Illegal Harms Call for Evidence suggested that predictive search functionalities play a role in increasing the discoverability of harmful and potentially illegal content on search services:
- a) Samaritans recommended that “autocomplete searches [are] turned off for harmful searches such as those relating to methods of harm and associated equipment.”<sup>808</sup>

---

<sup>806</sup> Volume 2: Chapter 6T Search Services paragraph 6T.21

<sup>807</sup> Ofcom, 2023. [Online content for use in the commission of fraud – accessibility via search services](#) . [accessed 22 September 2023].

<sup>808</sup> [Samaritans response](#) to 2022 Illegal Harms Call for Evidence.

- b) Antisemitism Policy Trust stated “search services have been found, through their systems, to direct people to hate material and racist content that is legal but can easily direct users to more extreme and illegal content when they follow search prompts.”<sup>809</sup> It also noted that Google’s Search autocomplete algorithm has been found to suggest antisemitic, racist and sexist content to users and that Microsoft Bing has been found to direct users to hateful searches via autocomplete.

22.13 There is also research that highlights the risks of predictive search:

- a) In 2019, TechCrunch commissioned a report by online safety startup AntiToxin on Microsoft Bing which found that Bing suggested keywords and images relating to CSAM.<sup>810</sup> When researchers input the keywords “Omegle Kids” (Omegle is an online service with chat, video and livestreaming functionality) into the search engine, Bing’s autocomplete suggested further terms which when searched surfaced illegal content.<sup>811</sup>
- b) The WeProtect Global Alliance notes that algorithms that suggest CSAM can have the effect of “encouraging or inspiring new offending, as well as increasing re-victimisation of those victims of abuse”.<sup>812</sup>

22.14 The evidence on the potential risk presented by predictive search functionalities focuses primarily on hateful search content (which, depending on the context, might amount to illegal content) and CSAM. However, we note the evidence from Samaritans relating to suicide and consider it reasonable to posit that predictive search could equally facilitate users encountering search content that is illegal content in other priority offence areas.

## Option

22.15 We have considered whether general search services that use predictive search functionalities should provide users with a means to easily report predictive search suggestions, and take appropriate steps to ensure that a reported suggestion is no longer recommended to users where it presents a clear and logical risk of users encountering priority illegal content.

## Effectiveness

22.16 We understand that it is the current industry standard practice across large general search services to enable user complaints in relation to predictive search suggestions. In some cases, this is accompanied by automated systems designed to prevent harmful results from being suggested and user controls to switch the functionality off.

22.17 We understand the following practices are in place:

- a) In its response to our 2022 Illegal Harms Call for Evidence, Google stated that it has “in-product reporting tools for many search features, such as autocomplete, and similar feedback mechanisms for other Search features, such as knowledge panels and featured

---

<sup>809</sup> [Antisemitism Policy Trust response](#) to 2022 Illegal Harms Call for Evidence, page 6.

<sup>810</sup> Microsoft removed the offending suggestions in response. Source: Constine, J., 2019. ‘[Microsoft Bing not only shows child sexual abuse, it suggests it](#)’, *TechCrunch*. 10 January [accessed 10 July 2023]. Subsequent references throughout.

<sup>811</sup> Constine, J., 2019.

<sup>812</sup> WeProtect Global Alliance, 2020. [Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse](#)- page 5. [accessed 13 September 2023]

snippets.”<sup>813</sup> It allows users to report violative autocomplete predictions.<sup>814</sup> It will remove autocomplete suggestions that violate its general or specific autocomplete policies, including where predictions contain dangerous, harassing, hateful or terrorist content.<sup>815</sup>

- b) Microsoft Bing similarly has the objective of preventing “inappropriate, offensive, or harmful predictions.”<sup>816</sup> It removes violative suggestions and enables users to turn search suggestions on or off. Within the reporting tool, “report a concern”, there is an option to report, “I have a concern similar to the above about Bing Image Creator, conversations, or content created by another AI powered feature.”

- 22.18 If a search service takes steps to remove reported predictive search suggestions that present a clear risk of directing users to illegal content, it would reduce the likelihood of other users being presented with these suggestions and potentially encountering illegal content via its service in future. This is particularly the case compared to circumstances where search predictions remain unmoderated. For example, a 2019 report by the Antisemitism Policy Trust and Community Security Trust (CST) found that once Google removed the “are Jews evil” search suggestion, 10% fewer related search requests asked whether Jews were evil in the 12 months following its removal compared to the 12 months prior.<sup>817</sup>
- 22.19 This suggests that the removal of suggestions deemed to present an illegal content risk would reduce the likelihood of users encountering illegal content in search results, because they would be less likely to search for it.
- 22.20 For users who are not actively searching for potentially illegal content, but are predisposed to searching for illegal content when prompted, this option would help prevent them from being suggested search terms that could lead to them encountering illegal content. For example, the evidence suggests that phrases and terms related to hate speech have been an issue in the past for some services.<sup>818</sup> We note, there may also be an incidental benefit for users who may be alarmed by the suggestion, but take no further action to search for illegal content – there is evidence of this occurring in relation to CSAM.<sup>819</sup>
- 22.21 For users that are actively searching for illegal content, this option would be unlikely to hinder their activities to a substantial degree, as they can still type in search requests to obtain the results they want.
- 22.22 However, we provisionally consider that the option would be effective at reducing the likelihood of users encountering search content that is priority illegal content.

---

<sup>813</sup> [Google response](#) to 2022 Illegal Harms Call for Evidence, page 24.

<sup>814</sup> Google. [Manage Google autocomplete predictions](#). [accessed 17 September 2023]

<sup>815</sup> Google Search Help. [How Google autocomplete predictions work](#). [accessed 11 July 2023]. Google Search Help. [Manage Google autocomplete predictions](#). [accessed 11 July 2023]

<sup>816</sup> Microsoft Support. [How Bing delivers search results](#). [accessed 11 July 2023] Subsequent references are to this website throughout

<sup>817</sup> Stephens-Davidowitz, S.S., 2019. [Hidden Hate: What Google searches tell us about antisemitism today](#), *Community Security Trust, Antisemitism Policy Trust*, page 19. [accessed 13 September 2023]

<sup>818</sup> Lapowsky, I., 2018. [Google Autocomplete still makes vile suggestions](#), *Wired*, 12 February [accessed 11 July 2023].

<sup>819</sup> Constine, J., 2019.

## Costs and risks

- 22.23 General search services are required to implement complaints and reporting systems under the Act to cover a wide range of topics. Therefore, any costs related to this option would be the incremental costs of adapting those systems to ensure the predictive search suggestions can be easily reported, and appropriate action taken. Generally, we would expect the extension of these systems to be relatively straightforward<sup>820</sup> and we would expect this option, relating to predictive search to take 20-40 days of software engineering time, along with an equal amount of non-software engineering time. This would approximately be equivalent to £9,000-£37,000 in one-off implementation costs.<sup>821</sup>
- 22.24 In addition to the implementation costs we would expect a service to incur ongoing costs. This would include the incremental maintenance costs of running the extended complaints system and the additional moderation costs that would be incurred when responding to complaints about predictive search. If the annual maintenance costs were 25% of the implementation cost, then this would be between £2,000-£9,000 per annum.<sup>822</sup>
- 22.25 The additional moderation costs to review the complaints received under this option are likely to vary depending on the size of the service. Larger services are likely to require a greater number of moderators as we would expect them to receive a larger number of user complaints. Services that use automatic moderation measures (such as Google<sup>823</sup>) may be able to make use of existing automated measures to process predictive search complaints. This may limit the increase in moderation costs if the service already has automated moderation functionality that can be adapted or is already able to handle these types of complaint.
- 22.26 Large general search services that utilise a predictive search functionality (Google and Bing) already make efforts to moderate their predictive search features to limit the likelihood that they do not suggest illegal or harmful content. This includes allowing users to report issues with predictive search, as described in paragraph [23.17] above.
- 22.27 Among smaller services, those that do not have predictive search functionality (such as Mojeek) and those that already have a complaints mechanism that can receive predictive search complaints (such as DuckDuckGo) would be expected to incur no or limited additional costs.
- 22.28 If recommended for smaller services, this option could reduce the potential for new services entering the search market, as they may need to incur the additional costs outlined above in order to compete effectively. In general, we expect the impact would be small in comparison to the other barriers that new search engines would face in entering the market and obtaining large reach.
- 22.29 Providers of general search services may remove predictive search functionalities to avoid the consequences of failing to properly moderate them, negatively impacting the user experience. We expect this to be a small risk since most services already have this measure in place and already have a complaints process that can be used for it. Similarly, there is

---

<sup>820</sup> For example, where a complaint system already exists, to allow an end-user to complain about contents of the auto suggest within a search term input box, the costs are incremental.

<sup>821</sup> Based on our labour cost assumptions set out in Annex 14.

<sup>822</sup> As described in Annex 14, we assume annual maintenance costs are 25% of the initial costs where we have no more specific information.

<sup>823</sup> Google. [Information Quality and Content Moderation](#) [accessed 14 September 2023]



likely to be only a small risk that services would over-moderate the predictive search algorithm which may cause the feature to lose functionality, as they have a commercial incentive for predictive search to be information for users.

- 22.30 Overall, we consider that the costs of this option are likely to be relatively low, however, we remain uncertain about the impact of this on smaller services, particularly as the cost of moderating complaints 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 those smaller services.
- 22.31 Given the actions already taken by Google and Bing we expect that they are likely to incur negligible or limited additional costs to implement this option. However, we recognize that applying this as an option would limit the ability for them to remove this function, which represents an additional cost if they had planned to remove this functionality.

## Rights impacts

### Freedom of expression

- 22.32 We believe that the impact of this option on freedom of expression would be limited to the search service provider, whose right to impart information to users in the form of predictive search suggestions would be restricted. However, we consider this restriction is proportionate to the overall reduction in the risk of harm to individuals in relation to illegal content that might have been encountered via the suggested search term. Reducing the risk of users encountering various kinds of illegal content, should contribute to the prevention of crime and the protection of physical and mental health, including the protection of children.
- 22.33 We do not consider there would be an impact on the user's right to receive information, as they are still able to freely input search requests and access information by means of the service. Similarly, there would be no impact on the right to impart information of persons who either operate or communicate via websites, as the website would remain operational and discoverable via the search service even where a predictive search suggestion that surfaces a URL was removed.

### Privacy

- 22.34 We believe the impact of this option on the right to privacy would be negligible. This option does not seek to alter the technical mechanism by which a predictive search algorithm functions, such as the reliance on a user's search history.
- 22.35 We acknowledge that user reports in relation to predictive search suggestions *might* generate new personal data or involve processing existing data for new purposes, if the service considered it appropriate to retain information about complainants (for example, for prioritisation purposes). Where complaints mechanisms involve personal data processing, services must comply with data protection laws. There would be no obligation on users to make complaints if they did not wish to. Finally, the risks are outweighed by the benefits: compliance with this option should lead to a reduction in illegal content harm.

## Provisional conclusion

- 22.36 This measure involves search services addressing concerns about predictive search suggestions. Doing this allows search services to reduce the risk that search predictions lead

users to illegal content they would not otherwise have encountered. We have set out above some specific examples for the kinds of harm this measure would help with. For example, it can help avoid leading people to view extreme and illegal hate content they might not otherwise have engaged with. And for fraud, it can make it more difficult for users to find prohibited fraud-related content. We consider that it is proportionate for the largest general search services given the potential benefits and relatively limited costs.

- 22.37 We understand that currently there are no large downstream general search services (using our proposed definition of more than 7 million UK users). However, in principle we see no reason to treat downstream general search services differently in relation to this measure.
- 22.38 For smaller general search services, we do not have evidence at this stage to conclude that the measure would be proportionate for them. This is because:
- a) Although the costs are likely to be relatively limited, they could still be material for a smaller service.
  - b) 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 no displacement risk from large services to small services, reducing the need to apply it to smaller services.
- 22.39 The reasoning we have set out above relating to the efficacy of this proposed measure relates to priority offences including CSAM, terrorism and other priority offences. We therefore propose to include this measure in our Codes for search services on terrorism, CSEA and other duties.
- 22.40 We propose to recommend that large general search services that use predictive search functionalities should:
- a) Develop and operate a mechanism that allows users to easily report predictive search suggestions which they consider to direct users towards illegal content;
  - b) Consider whether the wording of a reported predictive search suggestion presents a clear and logical risk of users encountering search content that is priority illegal content; and
  - c) If a risk is identified, take appropriate steps to ensure that the reported predictive search suggestion is not recommended to a user.
- 22.41 We consider that this measure would help large general search services meet their duty in [section 27(3)] to use proportionate systems and processes designed to minimise the risk of individuals encountering search content which is priority illegal content. The reduction in that risk would also help meet the duty in [section 27(2)], to effectively mitigate and manage the risks of harm to individuals, through users being less likely to search for harmful content.

## Measure to provide additional information to users about illegal content: CSAM content warnings

---

### Harms this measure seeks to address

- 22.42 As search services act as a gateway to the entire contents of the internet, it is possible that a user could use them to access, either inadvertently or deliberately, illegal content. As

outlined in the Register of Risks (Volume 2: 6T: Search services), general search services are identified as one of the most common methods of finding CSAM.

- 22.43 Evidence from the NCA suggests that general search services present a particularly acute risk of users encountering CSAM, and in particular that it can be found within three clicks on mainstream search engines.<sup>824</sup>
- 22.44 A qualitative study on the pathways for accessing CSAM online conducted interviews with 20 people who had viewed CSAM online and had been investigated by law enforcement. When asked about their initial exposure, two of the respondents reported that initial exposure occurred through intentional searches on search engines, and when asked about access methods, 13 respondents reported using search engines as a pathway to access CSAM.<sup>825</sup>
- 22.45 While our proposed measure in Chapter 15 in relation to deindexing of URLs known to contain CSAM would go some way to addressing these risks, there remains a residual risk that not all URLs with CSAM would be identified and search results could therefore contain CSAM.
- 22.46 Content warnings are designed to be surfaced when a user inputs a search query associated with CSAM and may act as friction in the user journey towards encountering illegal content via general search services. This can be a pop up containing a deterrent message, information on the potential offence, links to URLs for campaigns against the illegal content or support services or details on appropriate services to report potentially offending content.

## Option

- 22.47 In considering potential options to address these risks, we have considered the following possible measure:
- a) Large general search services should surface content warnings and support resources in response to user searches for CSAM.

## Effectiveness

- 22.48 The purpose of this option would be to provide deterrence and support to users inadvertently or intentionally attempting to access CSAM via search services. We anticipate that informing users of the illegality of CSAM may deter them, due to fear of facing legal action, from engaging with the search results or attempting to conduct such searches in the future.
- 22.49 Furthermore, the option would ask search services to provide links to resources designed to provide support and information to users which may help those that are purposefully seeking out such material to curb their behaviour.
- 22.50 We understand large general search services display content warnings in response to user searches for CSAM. For example:

---

<sup>824</sup>See principle 6 'example'. Source: Home Office, 2020. [Interim code of practice on online child sexual exploitation and abuse](#). [accessed 25 September 2023]

<sup>825</sup>Bailey, A. Allen, L., Stevens, E., Dervley, R., Findlater, D., & Wefers, S., 2022. [Pathways and Prevention for Indecent Images of Children Offending: A Qualitative Study](#). *Sexual Offending: Theory, Research, and Prevention*, 17. [accessed 14 September 2023]

- a) Google displays a deterrent message on searches for CSAM. It includes information on how to report CSAM to the IWF and a link to The Lucy Faithfull Foundation’s “Stop It Now!” campaign, which focuses on prevention of child sexual abuse and offers a broad range of support, including for those who are worried about their own thoughts or behaviour.
- b) Microsoft introduced measures in 2013 on Bing in collaboration with the Child Exploitation and Online Protection Centre (CEOP), which provides a list of keywords which when searched trigger a warning message.<sup>826</sup> It also presents a link to The Lucy Faithfull Foundation’s “Stop It Now!” campaign.

22.51 There is mixed evidence on the effectiveness of these content warnings as a means of directing users towards helplines and support. Some evidence points to the benefits of signposting to help services. For example, since 2015, 26,000 new users accessed the Stop It Now! self-help webpages as a result of splash pages directing them towards these resources.<sup>827</sup> A study conducted in 2014 on the effectiveness of the Stop It Now! helpline found 3.7% of callers had discovered the helpline through search engines. Those interviewed were supportive of the warning banners as a possible intervention to prevent the offending behaviour.<sup>828</sup> Similarly, since its launch in September 2021, the Finnish-based ReDirection program has been visited 80,000 times; most of these users accessed the webpages following intervention messages on dark-web search engines.<sup>829</sup> Of those users who went on to complete the ReDirection program, 77% said that their use of CSAM had reduced or stopped completely.<sup>830</sup>

22.52 By contrast, another qualitative study on the pathways for accessing CSAM online found that 14 out of 20 respondents reported not encountering any content warning messages online.<sup>831</sup> Of those who did encounter online content warning messages, some found these to be ineffective, however others suggested that warnings about the illegality of content and the consequences of viewing content could help prevent their viewing of CSAM. We recognise that in this case the sample size is small and that the respondents were not likely to be disposed to notice or adhere to the warnings, and therefore will not be a strong indication of the effect of warnings on search users more generally.

22.53 Content warnings may therefore not disrupt intentional searches in all cases. However, we do consider it would provide some useful friction for users intentionally searching for CSAM and may also be effective in mitigating the risk that users inadvertently access illegal content as a result of search requests. Further, given the correlation between viewing CSAM and going on to contact children for the purpose of committing further sexual offences, we consider that this measure would disrupt contact child sexual abuse, as well as the viewing

---

<sup>826</sup> Microsoft, 2013. [Microsoft and Google stand united to combat online child sexual abuse content](#). [accessed 17 September 2023]

<sup>827</sup> IWF. [URL Blocking and Filtering List](#). [accessed 29 September 2023]

<sup>828</sup> 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 14 September 2023]

<sup>829</sup> Protect Children, 2023, [‘Chat to a specialist’: Evaluation of an anonymous chat function of the ReDirection program](#), page 7. [accessed 25 September 2023]. Subsequent reference throughout.

<sup>830</sup> Protect Children, 2023, p.8.

<sup>831</sup> Bailey, A. Allen, L., Stevens, E., Dervley, R., Findlater, D., & Wefers, S., 2022. [Pathways and Prevention for Indecent Images of Children Offending: A Qualitative Study](#). *Sexual Offending: Theory, Research, and Prevention*, 17, 1-24. [accessed 14 September 2023]

of CSAM.<sup>832</sup> It may also benefit child victims of the offences documented in CSAM insofar as it may reduce the volume of CSAM encountered via search services.

## CSAM search terms

- 22.54 As part of this option, services would need to detect the nature of search terms entered by a user to deploy the warning. The evidence of current practice outlined in paragraph [23.50] suggests that it is possible for services to generate terms against which they consider it appropriate to provide warning information. In addition, we understand that expert third party organisations maintain keyword lists that they can share with services, which may be prepared in collaboration with law enforcement.
- 22.55 We are conscious that terms that may be used by offenders to search for CSAM vary in their specificity, and that it may not be appropriate to display a warning for each category. Broadly, we consider that terms used by offenders to search for CSAM fall into three categories:
- a) more obvious CSAM-specific layman terms, which clearly indicate that a person is seeking to encounter CSAM;
  - b) combinations of letters and symbols which are CSAM-specific and which are used to evade detection; and
  - c) seemingly innocuous terms known to generate CSAM results but which are not CSAM-specific.
- 22.56 We consider that presenting a warning in response to the latter category, seemingly innocuous terms which are not CSAM-specific, could have severe unintended consequences. In particular, this could inform the user that the term they have entered is CSAM related. We therefore provisionally consider that the risk of including such terms within this option is too high. We consider that it would be more appropriate for this option to be deployed in respect of more obvious CSAM-specific layman terms and combinations of letters and symbols which are CSAM-specific, neither of which would be used to generate non-CSAM results. Our provisional view is that these two categories clearly indicate that a user may be seeking to encounter CSAM. We acknowledge that this may have limitations, as it will not capture search requests using innocuous terms. However, we consider that it will provide effective in targeting offenders using other terms.
- 22.57 Regardless of whether a service chooses to develop their own list of terms, or use a third party list (or a combination of the two approaches), we consider that there are a number of principles that should be taken into account in developing or sourcing a list of CSAM search terms that fall within those two categories:
- a) The list should be developed by or sourced from a person with expertise in terms commonly used by offenders to search for CSAM online;
  - b) The list should be regularly updated with newly discovered terms, and to remove terms as relevant.
  - c) There should be arrangements in place to ensure that search terms are added to the list correctly. Where a list is sourced from a third party, the service should ensure that only

---

<sup>832</sup> A Protect Children study found that of respondents who had viewed CSAM, 37% had previously sought direct contact with children after viewing CSAM. Source: Insoll, T., Ovaska, A., and Vaaranen-Valkonen, N., 2021. [CSAM users in the dark web](#), *Protect Children*, page 40. [accessed 25 September 2023].

terms that fall within the categories identified in paragraph 23.55(a) and (b) above are used for the purposes of this measure<sup>833</sup>; and

- d) The list should be secured against unauthorised access, interference or exploitation by bad actors who may seek to obtain the list for the purposes of discovering (and possibly disseminating) terms which can be used to search for CSAM. This could include technical and non-technical measures, comprising of a mix of procedural, physical, personnel and technical controls.

## Cost and risks

- 22.58 For services that do not currently have this option in place, there would be initial costs to develop and implement it, in addition to ongoing costs to maintain and update the system to ensure it functions correctly. We expect that the upfront costs to develop a warning system would include the initial software development cost, and the development of a list of search terms related to CSAM in response to which the warning message would be shown.
- 22.59 In general, we expect services would make use of third parties with expertise to help develop a search term list, though our proposal leaves this to services to decide. As a result, costs could either come from purchasing external lists, developing their own list or a combination of both. As set out above, we expect that third party organisations already have keyword lists that they would share with services, and which are potentially available under the same arrangements as mentioned in the CSAM URL deindexing measure proposed in Chapter 15. This would be likely to reduce the implementation/running costs of this option.
- 22.60 Software costs would depend on whether services are making use of regularly updated third party lists. If so, they would need to ensure appropriate access controls to their system of the relevant third party. Other upfront software development costs could be material if services do not already have a system to provide warnings in response to search terms, however they could be much lower if they have an existing system to provide warnings/interstitials in other contexts.
- 22.61 We assume that the software development of applying this option would take between 170-310 days of software engineering time, as well as an equal amount of non-software engineer time depending on the complexity and existing functions of the system. We expect this to be equivalent to an implementation cost of up to be £80,000 - £290,000.<sup>834</sup> If the annual maintenance costs were 25% of the implementation cost, then this would be approximately £20,000- £70,000.<sup>835</sup> Ongoing running costs are likely to include updating the keyword list regularly and miscellaneous system maintenance costs.
- 22.62 If this measure is only applied to large search services, there is the potential for displacement where users move to smaller services to search using CSAM terms. However, we consider this risk is low because this measure is likely to be less effective for those users who may be more determined in searching for CSAM material. For example, we think it is

---

<sup>833</sup> We understand that many third party lists are developed for the purposes of content moderation, and as such contain terms that fall within the category identified in paragraph 23.55(c), which we do not propose to include as part of this measure for the reasons outlined in paragraph 23.56.

<sup>834</sup> Based on our labour cost assumptions set out in Annex 14.

<sup>835</sup> As described in Annex 14, we assume annual maintenance costs are 25% of the initial costs where we have no more specific information.

less likely that these users might move to a different search engine and instead we expect that these users are more likely to ignore any warning.

- 22.63 Both large general search services (Google and Microsoft) already have this measure in place. Therefore, we would not expect them to incur any additional costs unless they intended to remove this feature.
- 22.64 Moreover, search services operating in Australia that are subject to the eSafety Search Code would be subject to requirements that are similar in some respects to this option and would anyway need to take actions similar to those if we proposed to recommend this option.<sup>836</sup>

## Rights impacts

### Freedom of expression

- 22.65 We do not consider that this measure would have a material adverse effect on freedom of expression as users have no right to access CSAM. We recognise that it may discourage engagement with search content, however we consider that any potential interference of this nature is justified to prevent crime and protect health or morals. In particular, we note that the interstitials would only appear when a user enters search terms directly linked to CSAM, and therefore in circumstances where there is a high likelihood that they are seeking to encounter illegal content.

### Privacy

- 22.66 We don't consider there to be any impact on the right to privacy, as this option does not include that services retain information about searches conducted by individual users that might trigger the presentation of a content warning. Services which chose to retain information would need to do so in compliance with applicable privacy and data protection laws.

## Provisional conclusion

- 22.67 We outline above how this measure can be effective in helping some potential perpetrators curb their behaviour through education or fear of legal consequences. By showing potential perpetrators warning messages and providing links to resources and support services, the evidence suggests that some potential offenders will be less likely to access CSAM. In turn, this could mean less perpetrators going on to contact children for the purpose of committing further sexual offences. We therefore consider it is proportionate for large services to introduce this measure given the costs are likely to be relatively small compared to the measure's potential benefits in reducing the risk of harm. This is consistent with both of the existing large general search services (Google and Bing) currently undertaking this measure.
- 22.68 However, we are not proposing this measure for smaller services. This is because:

---

<sup>836</sup> 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 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, [Internet Search Engine Services Online Safety Code \(Class 1A and Class 1B Material\)](#), paragraphs 7(2)(g) and 7(2)(h). [Accessed 21 September]



- a) The costs associated with this measure are likely to be material for smaller services. Given the lower reach of these services, there are fewer potential perpetrators that are likely to be impacted by this measure. Alongside the mixed evidence of the effectiveness of this measure, we do not consider the benefits would necessarily outweigh the costs for those smaller services.
  - b) Our provisional view is that 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 and cease searching for CSAM, are more likely to ignore future warnings that move to a different, smaller search engine.
- 22.69 We are proposing to recommend as a part of our CSEA Code for search services that large general search services should employ appropriate means to detect and surface content warnings in response to user searches of which the wording clearly indicates that the user may be seeking to encounter CSAM and uses terms (or combinations of letters and symbols) that explicitly relate to CSAM.
- 22.70 Within this, services should:
- a) ensure that the warning:
    - i) informs users of the illegality of CSAM;
    - ii) provides links to resources and support services designed to help users refrain from committing CSEA offences freely available through a reputable organisation dedicated to tackling CSEA; and
    - iii) is prominently displayed in search results and is easy for users to understand;
  - b) develop and maintain a list of relevant search terms, either in-house or sourced from third party, in either case by a person with expertise in the terms commonly used to search for CSAM. Services should ensure that there are arrangements in place to ensure that:
    - iv) search terms are correctly added to the list and, where a list is sourced from a third party, that only search terms that meet the description in paragraph 23.69 are used for the purposes of this measure;
    - v) the list is regularly updated to add and remove relevant search terms as necessary; and
    - vi) the list is secured from unauthorised access, interference or exploitation.
- 22.71 We consider that this measure would be proportionate to help general search services meet their safety duty under section 27(3), under which search services must “minimise the risk of individuals encountering search content of priority illegal content or other illegal content.” By presenting content warnings to users on the risks of accessing illegal content via their site, the search service may contribute to minimising the risk of individual users encountering and engaging with search content that carries a high risk of including CSAM that has not yet been deindexed in line with the CSAM URL deindexing measure we propose to recommend in Chapter 15. Since users would be less likely to encounter CSAM content via the search service, this measure would also help meet the duty in [section 27(2)], to effectively mitigate and manage the risks of harm to individuals.

## Measure to provide additional context to users about illegal content: crisis prevention information

---

### Harms this measure seeks to address

- 22.72 As noted previously, general search services provide access to the contents of the entire internet which presents a risk that these services may be used to deliberately or inadvertently encounter illegal content. Search services are a gateway to information about suicide that exists online. Where that content intentionally encourages a person to end their life, or provides clear instructions on how to, this may amount to the priority offence of encouraging or assisting suicide.
- 22.73 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. As set out in our Register of Risks for Search services (Volume 2: 6T: Search services), there is evidence of the availability of a large volume of content relating to suicide online,<sup>837</sup> and of users accessing pro-suicide content via search services, some of which may meet the threshold of the priority offence.
- 22.74 There is some evidence to suggest that as suicidal intent increases, behaviour on search engines also changes, moving from periods of speculative browsing to specific and purposeful searches on methods of harm.<sup>838</sup> The risk of harm to those who are engaging in speculative, exploratory browsing sessions is that, alongside harmful content such as detailed discussions of methods of harm, they are likely to encounter content that intentionally encourages suicide.
- 22.75 We are also aware of research demonstrating the prevalence of this risk. A 2021 study investigating how search engines handle suicide queries examined the top 20 search results returned in response to queries related to suicide. The study found that 22% of Microsoft Bing URLs, 19% of DuckDuckGo URLs and 7% of Google Search URLs were “harmful”, that is, assessed by researchers to encourage, promote or facilitate suicide, or contain discussions of suicide methods. The researchers also looked specifically at search results encouraging suicide and found that this was the case for 10% of Microsoft Bing URLs, 8% of DuckDuckGo URLs and 4% of Google Search URLs.<sup>839</sup>
- 22.76 While the search results identified by the researchers as “encouraging suicide” may not have met the threshold of intention for the priority offence, these results nonetheless point to the potential risk of encountering such content for users either speculatively or purposefully browsing suicide content on search service. The risk of harm to users in that context is potentially very grave.

---

<sup>837</sup> Samaritans, 2020. [Understanding self-harm and suicide content online](#), page 3. [accessed 23 September 2023]

<sup>838</sup> Borge, O., Cosgrove, V., Cryst, E., Grossman, S., Perkins, S., & Van Meter, A., 2021. [How Search Engines Handle Suicide Queries](#). *Journal of Online Trust and Safety*, 1(1). [accessed 14/09/2023] Subsequent references throughout.

<sup>839</sup> Borge, O., et al, 2021.

## Option

22.77 In considering potential options to address these risks, we considered the option that all general search services should provide crisis prevention information for user searches for suicide.

## Effectiveness

22.78 If crisis prevention information is the first information that a user encounters in response to a search request relating to suicide, this may disrupt a search journey which could lead to illegal content that amounts to the offence of encouraging or assisting suicide.

22.79 Crisis prevention information can be surfaced in several ways, for example by ensuring crisis prevention services are prioritised in the search results or by providing crisis prevention information in an interstitial or banner.

22.80 We understand that suicide crisis prevention information is currently provided by several general search services:

- a) Google Search provides information in response to certain search requests relating to suicide and partners with crisis support services to display their information. For example, it provides the Samaritans helpline number, a facility to make a phone call via the mobile browser and a link to the official website of the Samaritans. In response to our call for evidence, Google said that this approach was a means of “connecting vulnerable users facing imminent harm with helpful and free resources immediately”.
- b) Microsoft Bing<sup>840</sup>, DuckDuckGo, Ecosia, AOL and Yahoo also present crisis support information in response to search requests including terms relating to suicide.

22.81 We understand that this current practice is broadly welcomed by charities operating in the mental health and suicide prevention space:

- a) In response to Google launching this functionality, Samaritans highlighted the importance of ensuring that “vulnerable and distressed people are steered towards safe spaces” given the large amounts of information that people can now access online.<sup>841</sup> The Samaritans have elsewhere suggested that search providers have a “corporate social responsibility” to promote sources of support in response to suicide-related search queries.<sup>842</sup>
- b) In response to our call for evidence for protection of children, Mental Health Innovations indicated that 2% (30-40 people) of its daily conversations on the Shout support service were referred via signposts on Google, and suggested that this demonstrates that “interventions such as this work to divert internet users” from potentially harmful searches.<sup>843</sup>

22.82 To implement this option services would provide links to freely-available information provided by reputable mental health charities and, given evidence that existing practice is successful at diverting users to helplines, that a helpline associated with such a charity is also

---

<sup>840</sup> Microsoft Support. [How Bing delivers search results](#). [accessed 12 July 2023]

<sup>841</sup> Samaritans, 2010, [Google and Samaritans: new search feature to help people looking online for information about suicide](#) [accessed 12 July 2023].

<sup>842</sup> Samaritans, 2013. [Samaritans and the online environment](#). [accessed 12 July 2023]

<sup>843</sup> MHIUK response to 2023 Ofcom Call for Evidence: Second phase of online safety regulation.

provided. This combination of information will ensure that the option operates effectively to prevent users encountering illegal content at a point of crisis. Services may choose to provide this information in such format they consider appropriate, provided that it is prominently displayed to users in the search results.

22.83 As part of this option, services would need to generate terms relevant to suicide and detect when they are entered by a user to deploy the information. The evidence of existing practice also suggests that it is possible for services to do so. Clearly, the fewer terms a service accurately identifies as potentially leading to illegal content that encourages suicide, the less effective this option will be. We provisionally consider that, as a minimum, it would be appropriate to expect services that services seek to cover search requests that fall within the following categories:

- a) General queries regarding suicide. The research referenced in paragraph [23.74] employed keyword research tools such as Google Trends and Semrush to identify popular suicide search terms, which included general searches such as “suicide” and “kill yourself”.<sup>844</sup> We recognise that this category is broad and may capture searches seeking help or pop culture references to suicide. However, it aligns with current practice of general search services (as outlined above) and we consider that including these more general terms may provide timely assistance for users, particularly at the earlier, speculative phase of browsing for suicide content; and
- b) Queries seeking specific, practical or instructive information regarding suicide methods, which may capture searches for instructions or resources about the experience of using one of those methods. 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 attempting suicide.<sup>845</sup> We therefore consider that providing crisis information at the point of conducting this more specific category of search request would be particularly effective at preventing users from encountering search content that encourages or assists suicide when in an extremely vulnerable state.

22.84 There may be other categories of search terms common among users experiencing thoughts of suicide, such as mood and anxiety symptoms or trauma and negative life events.<sup>846</sup> However, we would not expect services to include such categories of terms as the cost of developing a list of these terms may be considerably higher and it may result in the crisis prevention information being deployed in response to many searches with no connection to suicide.

## Cost and risks

22.85 For services that do not currently have this option in place, there would be initial costs to implement this option, and ongoing costs to maintain and update the system to ensure it operates correctly. If a service already has a system in place that can provide information in response to specific search terms, then we would expect the implementation costs to be

---

<sup>844</sup> Borge, O., et al, p.4, 2021.

<sup>845</sup> Moon KC, Van Meter AR, Kirschenbaum MA, Ali A, Kane JM, Birnbaum ML., 2021, [Internet Search Activity of Young People With Mood Disorders Who Are Hospitalized for Suicidal Thoughts and Behaviors: Qualitative Study of Google Search Activity](#). *JMIR Ment Health*. 8(10) [accessed 14 September 2023]

<sup>846</sup> Moon KC, et al, 2021.

moderate as this option would require a modification of an existing system to ensure that covers terms related to suicide.

- 22.86 We assume that to implement new functionality and capability of this nature would require approximately 150-310 days of software engineering time, along with an equal amount of non-software engineering time. This gives an estimated cost of approximately £70,000 - £290,000<sup>847</sup> including the cost to develop an interstitial displaying crisis prevention information. The total implementation cost would 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.
- 22.87 If the annual maintenance costs were 25% of the implementation cost, then this would be approximately £18,000-£70,000.<sup>848</sup> These running costs would likely cover system maintenance and updating the system to ensure it properly identifies search requests related to suicide.
- 22.88 However, this measure is already in place for both large general search services and several smaller search services. This suggests that in practice these costs are not excessive, at least for the large general search services.

## Rights impacts

### Freedom of expression

- 22.89 We consider that any impact on freedom of expression resulting from this measure would be limited and justified given the severe nature of the harm the measure is addressing. In theory, the option might have an impact on the freedom of expression rights of those who produce the substantial volumes of lawful content relating to suicide that exists online, and on the rights of users seeking to receive that content.
- 22.90 However, we think that any such impact would be limited given that users could still engage with the search results should they wish to do so. Moreover, to the extent that such an impact on freedom of expression exists, we consider that it is justified by the role the measure plays in contributing to the protection of health and prevention of crime.
- 22.91 Alongside freedom of expression, there may 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, as outlined above the presentation of this information would not prevent the user from engaging with the search results. We do not consider the effect would amount to an interference with freedom of association.

### Privacy

- 22.92 We don't consider there to be any impact on the right to privacy, as this option does not include that services retain information about searches conducted by individual users that might trigger the presentation of crisis prevention information.

---

<sup>847</sup> Based on our labour cost assumptions set out in Annex 14.

<sup>848</sup> As described in Annex 14, we assume annual maintenance costs are 25% of the initial costs where we have no more specific information.

## Provisional conclusion

- 22.93 As described above, we consider that there are significant benefits from the proposed measure in reducing the risk of harm related to suicide content. This is because it is likely to disrupt a user's search journey that could otherwise have led to illegal content related to suicide.
- 22.94 We consider that this measure is likely to be proportionate for large services, given these benefits and having regard to the level of costs outlined above. This is particularly the case given that Google and Microsoft Bing are already providing information of this type in response to search requests related to suicide.
- 22.95 At this stage we are not proposing to apply this measure to smaller services. The benefits are likely to be materially lower, as the lower reach of smaller services suggests that there are fewer journeys that are likely to be disrupted by the measure. Moreover, our analysis suggests that the costs smaller services would incur as a result of the measure could be material. As a result, it is unclear at this stage whether the measure would be proportionate for smaller services.
- 22.96 We also have some concerns that a measure of this type, with a relatively fixed cost of implementation, may be material for any new entrants that might be looking to enter the search market. This has also influenced our decision not to propose this measure for smaller services as we want to ensure that new entry and competition in the market for search services is not discouraged.
- 22.97 We note that a number of smaller services already voluntarily provide crisis information of this type (e.g., DuckDuckGo, Ecosia, AOL, Yahoo). We would encourage them to do so, notwithstanding the fact that we are not including them in the scope of this provision in codes at this time.
- 22.98 We are proposing to recommend as a part of our Code for other illegal content duties for search services that all large general search services should employ means to detect and provide crisis prevention information in response to search requests that contain:
- a) general queries regarding suicide; and
  - b) queries seeking specific, practical or instructive information or instructions regarding suicide methods.
- 22.99 Within this, they should ensure that the information:
- a) is prominently displayed to users in the search results;
  - b) is easy for users to understand; and
  - c) includes the following information:
    - i) a helpline associated with a reputable mental health or suicide prevention organisation; and
    - ii) link(s) to information and support that is freely available through a reputable mental health or suicide prevention organisation.
- 22.100 We consider that this measure would help general search services meet their duty in [section 27(3)] to use proportionate systems and processes designed to minimise the risk of individuals encountering search content which is priority illegal content. By reducing the risk of users encountering content that encourages or assists suicide, we also consider that this

measure helps general search services meet their safety duty under section 27(2), under which search services must “mitigate and manage the risks of harm to individuals.”

## Other measures considered

---

- 22.101 We have also considered whether to make recommendations in respect of the application of safe search, a feature which allows users to limit exposure to explicit and/or graphic content, as a measure to minimise the risk of users encountering illegal content. As part of this, we have considered whether to make recommendations regarding the default settings for safe search, which would require search engine providers to set a safe search feature to its strictest setting by default.
- 22.102 Safe search is a feature of several general search services which filters out results that are deemed explicit such as pornographic/sexual or violent content. Safe search features can have levels; for example, on Microsoft’s Bing the “Bing SafeSearch” feature can be set to strict, moderate, or off,<sup>849</sup> or can be opted in or out of such as with Google’s “SafeSearch”.<sup>850</sup> In some cases, a safe search feature is enabled by default, for example for children.
- 22.103 While safe search might capture content that is illegal, we have chosen not to consult on recommending it as a measure to comply with the illegal content safety duties at this stage. This is because we view safe search largely as a means of applying age-appropriate safety settings and a tool that is most appropriate for controlling the search content that children might encounter as a means of complying with the safety duties protecting children under section 29 of the Act. We will consider safe search measures when we consult on measures for the protection of children.
- 22.104 More broadly, in considering potential measures relating to the design and operation of search services, we had regard to the need for affording a higher level of protection for children than for adults, and to the needs of children of all ages in making use of search services, in line with the online safety objectives.<sup>851</sup> We believe that these objectives are generally better advanced through our protection of children code. That said, our proposed recommendations in this chapter and Chapter 15 (ACM for search services) that address the accessibility of CSAM via search services, including through deindexing and content warnings, will have a particular benefit for child victims of these offences by reducing the discovery of material which documents their abuse via search services.

---

<sup>849</sup> Microsoft Support. [Turn Bing SafeSearch on or off](#). [Accessed 04 July 2023].

<sup>850</sup> Google Search Help. [Filter or blur explicit results with SafeSearch](#). [accessed 05 July 2023].

<sup>851</sup> Paragraph 5(a)(v)/(vi) of Schedule 4 to the Act.



## 23. Cumulative Assessment of Proposed Measures

### What is this chapter about?

In the preceding chapters we assessed the impact of the measures we are proposing to include individually and explained why we think each of our proposals taken on its own is effective and proportionate. In this chapter we look at the cumulative impact of all our proposals taken together and assess whether, seen in the round, their impact would be proportionate. We focus in particular on the cumulative impact on small and micro businesses.

Our provisional conclusion is that not only are each of the measures seen on their own effective and proportionate, but that their cumulative impact would also be proportionate. In order to reach this conclusion, we have looked at the cumulative impact of the measures on three types of service: small low risk services; small services which are multi-risk or which pose a medium or high risk of a particular harm; and large services.

#### *Small low risk services*

All U2U and search services in scope of the Act will need to take some measures, even those provided by small and micro businesses that are low risk. For some services, these measures could require material changes. In order to ensure that the impact of the regulations is proportionate we have targeted the most onerous measures at the highest risk services. The assessment in this chapter indicates that by and large the impact of our proposals on small and low risk services should be low. Where measures in our Codes result in material costs for small and low risk services these costs result from explicit requirements of the Act rather than from decisions we have taken about how services should interpret the requirements of the Act.

#### *Small but risky services*

For those small and micro business that identify significant risks of illegal content in their risk assessments, we propose more demanding measures. These include additional governance measures, additional content or search moderation measures, and, in the case of services that pose a high risk of being used to disseminate CSAM potentially expensive measures such as hash matching.

The cumulative impact of these measures could be very significant and there is a possibility some small and micro businesses may even struggle to resource the recommendations we propose for them. However, on balance, we consider that the cumulative impact of our proposals is nonetheless proportionate given that we are targeting the costliest measures at high risk services.

#### *Large services*

For both U2U and search services, we are proposing more demanding measures for large services. This is partly because the benefits of large services taking measures tend to be greater due to their large user base. Also, they are likely to be able to access necessary resources to implement the measures.

### What input do we want from stakeholders?

- Do you agree that the overall burden of our measures on low risk small and micro businesses is proportionate?
- Do you agree that the overall burden is proportionate for those small and micro businesses that find they have significant risks of illegal content and for whom we propose to recommend more measures?
- We are applying more measures to large services. Do you agree that the overall burden on large services proportionate?

## Introduction

---

- 23.1 In the preceding sections we have assessed the impact of each of the measures we are proposing individually. In this section, we consider the cumulative impact of the measures taken together and explain why, seen in the round, we consider the package of measures to be proportionate. We consider the cumulative impact of our proposals on services provided by small and micro business, and then consider the implications of our proposals for large services.
- 23.2 The measures we recommend are summarised in the ‘Tear sheet’ published alongside this consultation. The tables included in this document show the measures we recommend in codes for U2U services and search services.
- 23.3 In those tables, each of the rows represents a different measure. The measures are grouped in the way we have discussed them in the different chapters of this consultation, which aligns with the way they are set out in the draft Codes.
- 23.4 Whether some of the measures are recommended for a particular service can depend on the size of the service and how risky it is. The different columns show different types of services. The columns are divided into two groups by size:
- a) Large services. As discussed further below, we propose to define a service as large where it has an average user base greater than 7 million per month in the UK, approximately equivalent to 10% of the UK population.
  - b) Smaller services. These are all services that are not large, and will include small and micro businesses.
- 23.5 We sub-divide each of these broad size categories into three:
- a) ‘Low risk’ relates to services that have assessed themselves as low risk for all kinds of harm in their risk assessment.
  - b) ‘Specific risk’ means a service has identified as medium or high risk for a specific kind of harm for which we propose a particular measure. Different harm-specific measures are recommended depending on which risk a service has identified. A service could have a single specific risk, or many specific risks. We are not currently proposing harm specific measures for all kinds of risk. The notes below the tables explain which kinds of risk different measures relate to.
  - c) ‘Multi risk’ means a service that faces significant risks for illegal offences in general. For such services, we propose additional measures that are aimed at illegal offences more generally, rather than being targeted at specific risks. As described in paragraph 11.41, our provisional view is to define a service as multi-risk where it has identified as

medium or high risk for at least two different kinds of harms from the 15 kinds of priority illegal offences set out in the Risk Assessment Guidance.<sup>852</sup>

- 23.6 Measures could be recommended for the same service from both the specific risk and multi risk columns, depending on the kinds of harms for which it is medium and high risk. In the extreme, if a service were medium or high risk for all kinds of harm, then all of the measures in the specific risk and multi-risk columns could apply to it.
- 23.7 Where the measure may apply to only some services to which a column relates, this is represented by a 'yes' in brackets. The additional conditions (aside from risk and the size of the service) affecting whether a measure is recommended are explained in notes after the tables.
- 23.8 The first column in the tables shows the measures that we recommend for a service if it were small and were low risk for all kinds of harm.

## Measures for small and micro businesses

---

- 23.9 For services that are not large (that is, those with fewer than 7 million monthly UK users), we propose fewer measures. Services provided by small businesses (fewer than 50 employees) and micro businesses (fewer than 10 employees) are likely to be in this category of services that are not large. This is because business would be likely to need more than 50 employees to provide a service to 7 million UK users, as discussed in paragraph 11.51. We assume below that no large services are provided by small and micro business.

## Measures recommended for low risk services provided by small and micro businesses

- 23.10 All U2U and search services in scope of the Act will need to take some measures to meet the significant and important new duties that the Act places on them. This includes services provided by small and micro businesses, even if those services have negligible risks.
- 23.11 The measures we propose recommending for all services, even if low risk, small and micro businesses, can be divided into two groups. The first group relate directly to specific duties in the Act. Examples include certain provisions in their terms of service or publicly available statements and for those provisions to be clear and accessible, and for all services to receive certain types of complaints. We have limited discretion over how this first group of measures should apply as the requirements in the Act are already very specific.
- 23.12 For some low risk, small and micro businesses, this first group of proposed measures could require material changes. This would be the case, for example, for those that do not currently have terms of service or a complaint handling function. Our impact assessment is not about the costs of the specific duties on services from the Act, as Ofcom is not making decisions about those specific duties. We are concerned with how our measures meet

---

<sup>852</sup> The 15 different kinds of illegal harms set out in Ofcom's draft risk assessment guidance are: Terrorism offences; Child Sexual Exploitation and Abuse (CSEA), including Grooming and Child Sexual Abuse Material (CSAM); Encouraging or assisting suicide (or attempted suicide) or serious Self Harm; Hate offences; Harassment, stalking, threats and abuse; Controlling or coercive behaviour (CCB); Drugs and psychoactive substances offences; Firearms and other weapons offences; Unlawful immigration and human trafficking; Sexual exploitation of adults; Extreme pornography offence; Intimate Image Abuse; Proceeds of crime offences; Fraud and Financial services offences; and Foreign Interference Offence (FIO).

those specific duties, and we consider our proposed measures set out a reasonable way of meeting those requirements, often giving services considerable flexibility in how they chose to do that.

- 23.13 Our impact assessment is focussed more on the second group of measures we propose. The second group of measures relate to the more general duties on services in the Act to protect users from illegal harms. We have more discretion over what this second group of measures should cover and who they should apply to. For U2U services, we propose only three additional measures for all services even if they are small and low risk:
- a) A named person is accountable to the most senior governance forum for compliance with illegal content duties, reporting and complaints duties;
  - b) Indicative timeframes for considering complaints should be sent to complainants; and
  - c) Accounts should be removed if there are reasonable grounds to infer they are run by or on behalf of a terrorist group or organisation proscribed by the UK Government (a “proscribed organisation”).
- 23.14 For small and low risk vertical search services, we only propose the first two additional measures above. We believe the cumulative impact of these two or three additional proposed measures on small and micro business that are low risk for all harms would be limited. For many such services, we expect the cumulative cost of these three measures to be less than a thousand pounds a year. This is assuming the named person does not need to do any more than they would otherwise have to do (the service would anyway have to, for example, undertake a risk assessment), the service receives very few complaints and it does not have any instances of accounts by proscribed organisations. We consider that small services would generally have the technical and financial capacity to undertake these measures.
- 23.15 We are considering the cumulative impact only of the measures we propose to recommend in Codes. As well as measures to address their main safety duties in the Act, all services will have to meet other requirements in the Act. This includes carrying out a risk assessment, which is a statutory requirement.
- 23.16 We anticipate that many small and micro businesses will be low risk for all kinds of harm. This is because the impact of many harms will often vary with the reach of the service as we set out in the Risk Assessment Guidance. Services provided by small and micro businesses will tend to have low reach, as services are likely to need higher numbers of employees if they have a large scale.
- 23.17 For general search services provided by small or micro businesses, we also propose a measure to ensure that CSAM URLs are deindexed from the search index. This is not linked to the service’s risk assessment, so it would apply even if a small general search services assessed itself to be low risk for all kinds of harm. We nevertheless consider this measure appropriate, partly because of the risk of users who are seeking CSAM material using such smaller services if they cannot access such material from large general search services. While this measure is fairly costly, we consider it necessary given the egregious nature of this harm. We anticipate the number of small general search services being very small in number, materially lower than the number of small U2U services.

## Measures recommended for multi-risk services provided by small and micro businesses

- 23.18 It is possible that some services provided by small and micro businesses will identify significant risks of illegal content in their risk assessments. In this case, we propose more demanding measures.
- 23.19 The measures we recommend for multi risk services are intended to help with all kinds of harm. These consist of additional governance and content moderation measures, and potentially collating safety metrics in recommender testing.
- 23.20 The total cost of these measures could be considerable. It will depend on many things, such as how complex the service is, how much illegal content users try to upload, the volume of complaints received, how easy the illegal harms are to deal with and the number of (human) moderators engaged. In many instances, we allow some flexibility in how services implement our recommendations, to ensure these are appropriate and proportionate to their circumstances, which allows services to implement in a way that is cost effective for them. Because of the variation by service, it is not possible to determine a precise estimate of the total costs of the measures applicable to multi-risk services. Nevertheless, it is clear that when combined with the other measures recommended for all services, the total cost for some small and micro businesses with significant risks of illegal harm could be considerable.
- 23.21 To give a sense of the scale of the costs, we are aware of a small service which needed to increase spending for online safety by several £100,000s per annum to deal with problematic content on its service relating to more than one harm area, where some of this material was illegal.<sup>853</sup> This cost is principally driven by content moderation costs, and the number of (human) moderators engaged. This suggests that the costs that some small and micro business will need to incur could be substantial.
- 23.22 Some small and micro businesses with significant risks may struggle to resource the recommendations we propose for them. Services would have the option of not following our Codes and describing how they have met their duties under the Act in another way, but they would still need to meet their safety duties in the Act. That such services need to incur costs if they are not already undertaking suitable measures is inevitable given the significant and important new safety duties that the Act places on them. It is even possible that some such services may cease to operate in the UK, or cease to operate at all if the UK is an important market for them.
- 23.23 Even if some small services with significant risks would cease to operate in the UK, this does not necessarily mean that the measures are disproportionate. While there is likely to be a loss to society from any services ceasing to operate, and from potential entrant services choosing not to start to operate, this needs to be considered in the context of these services posing significant risks to society from illegal conduct and content.
- 23.24 Whilst the cumulative impact of the measures under consideration could be significant, our provisional view is that it would be proportionate. This is because we are targeting them at services that pose a material risk of causing significant harm to people in the UK. Given that

---

<sup>853</sup> This is based on the increase in the number of content moderators that BitChute plans to put in place. This will increase to 21 content moderators, which we have used this to estimate the costs above.

we think the measures would be effective in tackling this harm, we consider that the benefits justify the costs of the measures and the impact they would have on business.

## Measures recommended for specific risks for services provided by small and micro businesses

- 23.25 In addition to the measures to address harms in general, we also propose recommending more measures for services where they assess as high risk for CSEA, and have specific functionalities.
- 23.26 Some of these measures would be costly. For example, for some U2U services that identify a high risk of grooming, we propose to recommend a range of measures to change their defaults which could be costly, including in terms of reduced engagement and revenue. For any small service with more than 700,000 UK users that has identified as high risk for image-based CSAM, we propose to recommend adopting CSAM hash matching. For file-sharing services that identify as high risk for CSAM, we propose they adopt CSAM hash matching even if they only have 70,000 UK users, because this kind of service is particularly risky given the significant role they play in the circulation of CSAM.
- 23.27 These measures can entail significant costs and it is possible that some small businesses that have high risks of CSEA may struggle to resource all of the recommendations we propose for them. Nevertheless, our provisional view is that the cumulative impact of these measures is still proportionate for these smaller businesses. This is because the potential harms are very considerable for online child sexual exploitation and abuse online, and we see reducing this as a strategic priority. We welcome views on this.

## More onerous measures for large services

---

### Measures recommended for large services that are low risk

- 23.28 We are proposing more demanding measures for large services, for both U2U and search. This is the case even if a large service assesses as low risk for all kinds of harm in its risk assessment. For large services that are low risk, we propose various extra governance and general content moderation measures, as set out in the tables in the Tear sheet.
- 23.29 For each of these measures, we have explained in the relevant chapter why we propose recommending it for large services even if they assess as low risk. This is often related to the risk of a failure in governance and content moderation potentially affecting a large number of users, and so potentially having a significant adverse impact. As the nature of illegal harm can change over time, having suitable governance and content moderation measures in place can help manage new and escalating risks quickly and effectively. Moreover, large services are likely to be more complex because of their size, and there is a greater possibility that some risks have not been examined properly, meaning that governance measures in particular are more important.
- 23.30 We propose making an exception for vertical search services and not to recommend some of the governance and content moderation measures for such services just because they are large. We are not aware of evidence of such services showing illegal content and by their nature vertical search services are unlikely to have content that is as rapidly changing as U2U services, and the search results are more under their control than for U2U

content.<sup>854</sup> Any benefits of recommending such measures to these services would therefore be low and we do not consider it proportionate. For many of these measures, the costs of implementing them are likely to be fairly low if a service is low risk. We therefore do not anticipate these measures being overly burdensome for large services that are low risk for all harms.

- 23.31 Our provisional view is that these additional governance and content moderation measures are proportionate for low risk services that are large. There are particular reasons for wanting large services to undertake these measures to keep users safe, the costs are unlikely to be that high if the services are low risk, and large services are likely to have the resources to undertake these measures.
- 23.32 In practice, we anticipate that large services will generally identify themselves as risky for at least some harms, as their large reach tends to increase the impact of any illegal content. We therefore do not envisage there being many large services that are low risk for all kinds of illegal harm.
- 23.33 For large general search services, we propose recommending various measures independent of those services' risk assessments. These include measures relating to predictive search, warning messages for CSAM related searches and crisis prevention information for suicide related searches. We do not consider it necessary to make these measures contingent on the service's risk assessment, as we consider that all such services would generally be risky for the relevant harms given their wide reach. We therefore do not regard these measures as applying to low risk large general search services, because there would be no such services. There are currently only two large general search services, namely Google and Bing.

## Measures recommended for multi-risk large services

- 23.34 For large services that are multi-risk, we anticipate that the above extra governance and content moderation measures would be more costly than for low risk large services. We also propose an additional governance measure for such services. This is to have an internal monitoring and assurance function to independently assess the effectiveness of the mitigations of illegal harms.
- 23.35 For U2U services that have recommender systems to determine the relative ranking of content and undertake on-platform tests, we also propose such services collect safety metrics. This will allow them to assess whether the changes are likely to increase user exposure to illegal content.
- 23.36 We envisage the internal monitoring and assurance function in particular to be a costly measure. But we consider these additional one or two measures (depending on whether the service is U2U or search) to be proportionate for large services with significant risks. This is partly because the benefits of large services taking measures tend to be greater due to their large user base. Also, large services are likely to be able to access necessary resources to implement the measures.
- 23.37 We expect large services, those with over 7 million monthly UK users, to have the capacity to undertake the proposed measures. As discussed from paragraph 11.55, we have also

---

<sup>854</sup> See Volume 2: Chapter 6T (Search services), paragraph 6T.21(b) for why we consider vertical search services to be low risk.



considered some possible ways in which the definition of large services could be narrowed, such as relating to employees or financial numbers, to reduce the possibility of large service having limited resources. On balance, we propose to keep the definition simple and relate only to user reach. In the context of the range of measures that we propose to recommend for large services, we welcome views on whether we should consider supplementing our definition of large.

## **Measures recommended for specific risks for large services**

- 23.38 We propose various other measures for large U2U services if they identify as medium or high risk for specific kinds of harm, with some of these measures contingent on having specific functionalities. As set out in the Tear sheet, these include measures relating to dedicated reporting channels, fraud (stolen credentials) standard keyword search, enabling a user to block or mute other users, enabling users to disable comments, and conditions relating to notable user verification schemes.
- 23.39 We do not always have detailed information on the likely costs of these measures, which is why we are not recommending for smaller services at this time. Nevertheless, we expect that the benefits are likely to justify the costs for large services. This is partly because the benefits of applying these measures to large services tend to be larger given their greater large user base. We believe large services will have the resources to bear the cost of these measures. This view is supported by the fact that several large services already have comparable processes in place for many of these measures.
- 23.40 Our provisional view is that the cumulative impact of these measures to address specific risk, on top of the multi-risk measures, is proportionate for large services.

# 24. 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, including requiring us to have regard to the risk of harm to citizens presented by content on regulated services.

In this chapter we outline 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 proposed recommendations for our illegal content Codes of Practice meet these requirements. We provide further information regarding Ofcom's duties relating to the preparation of our Codes in our Legal Framework (Annex 12).

## What input do we want from stakeholders?

- Do you agree that Ofcom's proposed recommendations for the Codes are appropriate in the light of the matters to which Ofcom must have regard? If not, why not?

## Background

---

- 24.1 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.
- 24.2 In Chapters 12 to 22, we set out our proposed recommendations; an overview of these recommendations can be found in Chapter 11, and our cumulative assessment of the measures can be found in Chapter 23. The draft measures themselves can be found in full in Annex 7 (U2U) and Annex 8 (Search). We provide further information regarding Ofcom's duties relating to the preparation of our Codes in our Legal Framework (Annex 12).
- 24.3 We consider that our proposals 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.

## Appropriateness and principles

---

- 24.4 As required by section 3 of the Communications Act 2003, in making the proposed recommendations Ofcom has had regard to the matters set out below and to the risk of harm to citizens presented by content on regulated services.
- 24.5 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 has set out in the Consultation our reasons for proposing to apply some Code recommendations to services of different kinds, sizes and capacities.
- 24.6 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.

- a) Ofcom has clearly identified in our draft Codes which measures apply to what types and sizes of services, for the reasons given in each relevant section of the Consultation.

**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.

- b) Having regard to the need for it to be clear to providers of regulated services how they may comply with their duties dealt with in this Consultation, Ofcom has aimed to be as clear and detailed as possible in our draft Codes, consistent with acting proportionately.

**Paragraph 2(c):** the measures described in the code of practice must be proportionate and technically feasible: measures that are proportionate or technically feasible for providers of a certain size or capacity, or for services of a certain kind or size, may not be proportionate or technically feasible for providers of a different size or capacity or for services of a different kind or size;

- c) Ofcom is proposing to recommend measures many of which we know to be in widespread use in the sector. Ofcom has clearly identified in our draft Codes which measures apply to what types and sizes of services, for the reasons given in each relevant section of the Consultation.

**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 [89] of the risk of harm presented by services of that kind or size.

- d) Ofcom has identified in our reasoning the harms which our proposed recommendations would address, and explained why we consider each proposed measure is proportionate in the light of those harms. As required by section 3(4A)(b)(ii) of the 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 draft Register of Risk. Where appropriate, Ofcom has clearly identified in our draft Codes which measures would apply to what types and sizes of services, for the reasons given in each relevant section of the Consultation.

- 24.7 Having had regard to the desirability of promoting the use by providers of regulated services of technologies which are designed to reduce the risk of harm to citizens presented by content on regulated services, and to the seriousness of the harms concerned, Ofcom has, in particular, recommended the use of certain kinds of technologies where proportionate to the risk of harm from CSAM and fraud (see, for example, Chapters 14 and 15). Having regard to the desirability of encouraging investment and innovation in the markets for regulated services and these technologies, our proposals do not recommend specific technologies or the use of specific inputs, in order to secure that services can act in accordance with our recommendations using any appropriate technology or input.

## Ofcom's general duties and the online safety objectives

---

### U2U services

24.8 As required by paragraph 3 of Schedule 4 to the Online Safety Act, Ofcom has also ensured that the proposed recommendations are compatible with the pursuit of the applicable online safety objectives for U2U services as follows:

**Paragraph 4(a)(i):** a service should be designed and operated in such a way that the systems and processes for regulatory compliance and risk management are effective and proportionate to the kind and size of service.

- a) In Chapter 8 (Governance and accountability), Ofcom has set out the governance measures which we propose to recommend having regard, amongst other things, to the kind and size of service. We consider these to be compatible with this objective.

**Paragraph 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.

- b) 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 Chapter 8 (Governance and accountability), Chapter 12 (U2U Content Moderation), Chapter 14 (U2U Automated Content Moderation), Chapter 16 (Complaints and Reporting), and Chapter 20 (Enhanced User Controls), Ofcom has set out the systems and processes measures which we propose to 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

- c) In Chapter 17 (ToS) we are consulting on a proposed recommendation which we consider would be compatible with this objective.

**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.

- d) In Chapter 16 (Complaints and Reporting), and Chapter 18 (Default settings and support for child users) we are consulting on proposed recommendations which we consider would be compatible with this objective.

**Paragraph 4(a)(vi):** a service should be designed and operated in such a way that the service provides a higher standard of protection for children than for adults.

- e) Having regard to the need for a higher level of protection for children than for adults, in Chapter 18 (Default settings and support for child users) we are consulting on proposed recommendations which we consider would be compatible with this objective.

**Paragraph 4(a)(vii):** a service should be designed and operated in such a way that the different needs of children at different ages are taken into account.

- f) In Chapter 17 (ToS), and Chapter 18 (Default settings and support for child users) we set out how we have had regard to the different needs of children at different ages. We are consulting on proposed 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.

- g) In Chapter 21 (User Access) we set out why we do not consider it appropriate to restrict access to services generally by adults. We explain the measures on which we are consulting to limit the activities of proscribed organisations. In Chapter 20 (Enhanced 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.

- h) In Chapter 18 (Default settings and support for child users) we are consulting on proposed recommendations which we consider would be compatible with this objective, and explain how we have taken into account use of the service by, and impact on, children in different age groups.

**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.

- i) All our recommendations seek to protect users from harm. In particular, in Chapter 8 (Governance and Accountability), Chapter 12 (U2U Content Moderation), Chapter 14 (U2U Automated Content Moderation), Chapter 16 (Complaints and Reporting), Chapter 18 (Default settings and support for child users), and Chapter 19 (Recommender System Testing), we are consulting on proposed recommendations which we consider would be compatible with this objective.

24.9 We are not at this stage consulting on measures relating to paragraph 4(a)(v) given it is specific to Category 1 services only.

## Search services

24.10 As required by paragraph 3 of Schedule 4 to the Online Safety Act, Ofcom has ensured that the proposed recommendations are compatible with the pursuit of the applicable online safety objectives for search services as follows:

**Paragraph 5(a)(i):** a service should be designed and operated in such a way that the systems and processes for regulatory compliance and risk management are effective and proportionate to the kind and size of service.

- a) In Chapter 8 (Governance and accountability), Ofcom has set out the governance measures which we propose to recommend having regard, amongst other things, to the kind and size of service. We consider these to be compatible with this objective.

**Paragraph 5(a)(ii):** a service should be designed and operated in such a way that the systems and processes are appropriate to deal with the number of users of the service and its user base.

- b) In Chapter 8 (Governance and accountability), Chapter 13 (Search Moderation), Chapter 15 (Search Automated Content Moderation), and Chapter 22 (Search Service Design and User Support), Ofcom has set out the systems and processes measures which we propose to 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 23 and 25.

- c) In Chapter 17 (PaS) we are consulting on a proposed recommendation which we consider would be compatible with this objective. In making this recommendation, our duty to have regard to the extent to which providers of regulated services demonstrate, in a way that is transparent and accountable, that they are complying with their duties set out in the Act, is relevant.

**Paragraph 5(a)(iv):** a service should be designed and operated in such a way that there are adequate systems and processes to support United Kingdom users.

- d) In *Chapter 13 (Search Moderation)*, *Chapter 15 (Search Automated Content Moderation)*, and *Chapter 22 (Search Service Design and User Support)* we are consulting on a proposed recommendation which we consider would be compatible with this objective.

**Paragraph 5(a)(v):** a service should be designed and operated in such a way that the service provides a higher standard of protection for children than for adults.

- e) Having had careful regard to the need for a higher level of protection for children than for adults, in *Chapter 15 (Search Automated Content Moderation)* we are consulting on proposed recommendations which we consider would be compatible with this objective. Outside of these specific recommendations, for the reasons set out in *[paragraph X of search]* we consider that this objective is better advanced via the Codes on protection of children than the Codes in relation to illegal content.

**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.

- f) In Chapter 16 (Complaints and Reporting), Chapter 17 (PaS) we set out how we have had regard to the different needs of children at different ages. We consider, outside of

these specific recommendations, that this objective is better advanced via the Codes on protection of children than the Codes in relation to illegal content.

**Paragraph 5(b):** a service should be assessed to understand its use by, and impact on, children in different age groups.

- g) We have had regard to the needs of children of all ages, but consider that this objective is better advanced via the Codes on protection of children than the Codes in relation to illegal content. We consider our recommendations in relation to illegal content 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

- h) In Chapter 8 (Governance and accountability), Chapter 15 (Search Automated Content Moderation), and Chapter 22 (Search Service Design and User Support) we are consulting on proposed recommendations which we consider would be compatible with this objective.

## Content of Codes of Practice

---

### U2U services

- 24.11 Codes of practice that describe measures recommended for the purpose of compliance with a duty set out in section 9(2) or (3) (illegal content) must include measures in each of the areas of a service listed in section 9(4). This provision applies to the extent that inclusion of the measures in question is consistent with:
- a) Ofcom's duty to consider the appropriateness of provisions of the code of practice to different kinds and sizes of Part 3 services and to providers of differing sizes and capacities;
  - b) the principle that the measures described in the code of practice must be proportionate and technically feasible; and
  - c) the principle that the measures described in the code of practice that apply in relation to Part 3 services of various kinds and sizes must be proportionate to OFCOM's assessment (under [section 89]) of the risk of harm presented by services of that kind or size.
- 24.12 Ofcom has made proposals for U2U services in each of the areas of a service listed in section 9(4) as follows:
- a) regulatory compliance and risk management arrangements – see Chapter 8 (Governance and accountability),
  - b) design of functionalities, algorithms and other features – see Chapter 18 (Default settings and support for child users), and Chapter 19 (Recommender System Testing),]



- c) policies on terms of use – see Chapter 17 (ToS), and Chapter 12 (U2U 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 21 (User Access),
- e) content moderation, including taking down content – see Chapter 12 (U2U Content Moderation), and Chapter 14 (U2U Automated Content Moderation),
- f) functionalities allowing users to control the content they encounter – see Chapter 18 (Default settings and support for child users), Chapter 20 (Enhanced User Controls),
- g) user support measures – see Chapter 18 (Default settings and support for child users),
- h) staff policies and practices – see Chapter 8 (Governance and accountability), and Chapter 12 (U2U Content Moderation).

## Search services

- 24.13 Codes of practice that describe measures recommended for the purpose of compliance with a duty set out in section 23(2) or (3) (illegal content) must include measures in each of the areas of a service listed in section 23(4). This provision applies to the extent that inclusion of the measures in question is consistent with:
- a) Ofcom’s duty to consider the appropriateness of provisions of the code of practice to different kinds and sizes of Part 3 services and to providers of differing sizes and capacities;
  - b) the principle that the measures described in the code of practice must be proportionate and technically feasible; and
  - c) the principle that the measures described in the code of practice that apply in relation to Part 3 services of various kinds and sizes must be proportionate to OFCOM’s assessment (under [section 89]) of the risk of harm presented by services of that kind or size.
- 24.14 Ofcom has made proposals for search services in the following areas of a service listed in section 23(4) as follows:
- a) regulatory compliance and risk management arrangements – see Chapter 8 (Governance and accountability),
  - b) design of functionalities, algorithms and other features relating to the search engine – see Chapter 16 (Complaints and Reporting), *Chapter 15 (Search Automated Content Moderation)*, and *Chapter 22 (Search Service Design and User Support)*,
  - c) user support measures – see *Chapter 22 (Search Service Design and User Support)*
  - d) staff policies and practices – see Chapter 8 (Governance and accountability), and *Chapter 13 (Search Moderation)*.

For the reasons set out in the relevant sections, Ofcom did not consider it appropriate or proportionate to make proposals for search services in relation to illegal content in relation to one area of a service listed in section 23(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.