Protecting people from illegal harms online

Summary of each chapter

Published 9 November 2023

Contents

Section

Overview

This document sets out a high-level summary of each chapter of our illegal harms consultation to help stakeholders navigate and engage with our consultation document. The full detail of our proposals and the underlying rationale, as well as detailed consultation questions, are set out in the full document. This is the first of several consultations we will be publishing under the Online Safety Act. Our full regulatory roadmap and strategy is available on our website.

We are consulting publicly on the proposals set out in full in this consultation document. We welcome comments from stakeholders in response to our proposals, including any further evidence and supporting information to inform our final decisions. You can find out more about how to respond to our consultation, our consultation principles and our detailed consultation questions in annexes 1-4.

Please respond to our consultation by completing and submitting the consultation form to IHconsultation@ofcom.org.uk. This consultation closes on 23 February 2024.

Index

Volume 4 – What should services do to mitigate the risk of online harms?

Volume 5 – How to judge whether content is illegal or not

Volume 6 – Our information powers, enforcement powers and approach to supervision

Volume 1: Background to the new Online Safety regime (introduction, illegal content duties and offences, and overview of regulated services)

1. Introduction

This section provides a high-level introduction to this consultation (illegal harms consultation) on putting into effect the illegal content duties and our enforcement powers under the Online Safety Act 2023 (the 'Act'). It outlines the broad scope of the consultation, our duties and safety functions, and explains how to use and navigate this document.

2. Illegal content duties and overview of relevant offences

This chapter summarises the main duties the Act creates. The Act gives online services a range of duties. The main ones relating to illegal content are for services to assess the risk of harm arising from illegal content (for a user to user (U2U) service) or activity on their service, and take proportionate steps to manage and mitigate those risks.

The Act lists over 130 'priority offences'. U2U services will need to act to prevent users encountering content amounting to one of these offences and search services will need to minimise the risks of users encountering content that amounts to one of these offences. We have grouped these offences into broad groups, such as terrorism, hate, child sexual exploitation and abuse, sexual exploitation of adults, unlawful immigration and human trafficking.

Services also have duties to swiftly take down certain types of non-priority illegal content.

3. Overview of regulated services

What is this chapter about?

This chapter explains which types of services are in scope of the Act. The Act places new legal requirements on providers of the following three types of internet service: services that allow 'user-to-user' interactions or 'user-generated content'; search services; and providers of pornographic content.

The duties in the Act apply to services with links to the UK regardless of where in the world they are based. The number of online services subject to regulation could total more than 100,000 and range from some of the largest tech companies in the world to very small services. Services in scope of the Act come from a diverse range of sectors, including, but not limited to, social media, dating, gaming and adult services.

The online space is one of rapid innovation. We know that new types of U2U and search services will emerge, a good recent example being developments in generative AI.

This has a number of implications for our work:

- Firstly, we will flex our expectations depending on the type of service we are dealing with – we will not expect the same of a small low risk service as we do of the largest or riskiest services.
- Secondly, we will need to adapt our approach and expectations over time to reflect the emergence of new technologies and types of U2U or search services. We will scan the horizon for new developments and, when necessary, we will update our codes to reflect the emergence of new risks and new options for mitigating risks. As we explain below, we will also expect services to monitor the emergence of new risks.
- Thirdly, as described in our background section[1] we will need to use a combination of different regulatory levers to achieve our goals and to use different levers to influence different types of service. For example, sometimes we will seek to drive change by: setting expectations in our codes of practice; taking enforcement action against services which are not complying with the regulations; using our research and our transparency reporting powers to shine a light on what services are doing to tackle online harms and generating reputational incentives for them to make improvements; and engaging with services and discussing with them where we consider they should be doing more to improve user safety.

[1] See the document titled, "Ofcom's approach to implementing the Online Safety Act" for further detail.

Volume 2: The causes and impacts of online harm

4. Introduction to volume 2

In this volume, we set out our understanding of the causes and impacts of online harm. We explain how we have compiled our evidence base for our sector-wide risk assessment, set out our key findings and detail the analysis of our sector-wide risk assessment in our 'Register of Risks'.

5. Evidence and methodology for conducting our risk assessment

This chapter explains how we have conducted the analysis of our sector-wide risk assessment (presented in the 'Register of Risks'), including the evidence used, the offences considered, and the risk factors analysed. This chapter also looks at the considerations involved in assessing the risk of harm to individuals.

6. The causes and impacts of online harms: Ofcom's register of risk

This chapter presents our assessment of the causes and impacts of illegal online harms based on the evidence that we have gathered over the past three years. The analysis we set out here forms part of our duty under the Act to assess the factors that can cause a risk of harm to individuals on a service. We expect services to have reference to it when they carry out their own risk assessments. Our assessment focuses on the over 130 priority offences defined in the Act. For ease of navigation, we have grouped these into 15 broad kinds of illegal harm. These include illegal harms such as: Child Sexual Exploitation and Abuse (CSEA), terrorism, fraud and hate speech, as well as the newly created Foreign Interference Offence, which addresses malicious online activity conducted by foreign powers (for example, state-sponsored disinformation campaigns). We summarise the findings of this assessment below and set out the detailed analysis in the body of this volume.

The illegal harms we have looked at are widespread and, in many cases, growing in prevalence. For example, 87% of adult internet users report having encountered a scam or fraud online and 25% of these people have lost money as a result. Almost a fifth of children experience sexual solicitation from adults they have chatted with online, and there was a 707% increase in the number of Uniform Resource Locator (URLs) which contain Child Sexual Abuse Material (CSAM) reported to the IWF between 2014 and 2021.

Online harms affect a large proportion of people in the UK and a wide cross-section of society. 63% of UK internet users say they have encountered potentially harmful content online in the past four weeks. However, children and people with certain protected characteristics are most likely to be affected. For example, 16% of minority ethnic internet users have encountered 'hateful, offensive or discriminatory content', compared to 11% of all internet users. Similarly, studies have shown that women are five times more likely to be victims of intimate image abuse. The more protected characteristics someone has, the more at risk of harm they are from priority illegal harms in the Act.

The impact of the harms we have looked at can be extremely severe. It is not limited to the online world but can also profoundly affect people's lives offline. A particularly clear example of this is online grooming, which can result in contact sexual abuse and cause lifelong negative psychological impacts including, loss of confidence, aggression, feelings of self-blame and lack of personal trust, as well as an increased risk of self-harm. In most cases the harms we have looked at primarily affect the individual experiencing them. However, in some cases they have a wider impact on society as a whole. For instance, state-sponsored disinformation campaigns can erode trust in the democratic process. All this underlines the need for the new legislation and shows that, while many services have made significant investments in tackling online harm in recent years, these have not yet been sufficient.

The kinds of illegal harm we have looked at occur on services of all types. Services as diverse as social media services, dating services, marketplaces and listings services, search services, adult services, and file-storage and file-sharing services are all used to disseminate some of the types of harmful content we have looked at in this volume. Bad actors use both large and small services to spread illegal content, although the way in which they use large services sometimes differs from the way in which they use small services. For example, terrorists often use large services to disseminate propaganda to large audiences, but often use small services for more covert activities such as recruitment, planning and fundraising. Related to this, offenders often rely on multiple different types of service to commit or facilitate the offences covered by the Act. For instance, both fraudsters and perpetrators of grooming will often contact potential victims on public forums and then seek to move them onto private, encrypted services. This means that action to tackle online harms cannot focus exclusively on a small subset of services and cannot be targeted exclusively at the largest services. Rather, it needs to address a broad range of service types including both large services and the long tail of smaller services in scope of the Act.

Although a very wide range of service types pose risks of the priority illegal harms in the Act, certain service types appear to play a particularly prominent role in the spread of priority illegal content. In particular, our analysis suggests that file-storage and file-sharing services and adult services pose a particularly high risk of disseminating CSAM, and social media services play a role in the spread of an especially broad range of illegal harms. Similarly, certain 'functionalities' stand out as posing particular risks:

- **End-to-end encryption**: Offenders often use end-to-end encrypted services to evade detection. For example end-to-end encryption can enable perpetrators to circulate CSAM, engage in fraud, and spread terrorist content with a reduced risk of detection.
- **Pseudonymity and anonymity**: There is some evidence that pseudonymity (where a person's identity is hidden from others through the use of aliases) and anonymity can embolden offenders to engage in a number of harmful behaviour with reduced fear of the consequences. For example, while the evidence is contested, some studies suggest that pseudonymity and anonymity can embolden people to commit hate speech. At the same time, cases of harassment and stalking often involve perpetrators creating multiple fake user profiles to contact individuals against their will and to circumvent blocking and moderation.
- **Livestreaming**: There are many examples of terrorists livestreaming attacks. This can in turn incite further violence. The use of livestreaming remains a persistent feature of far-right lone attackers, many of whom directly reference and copy aspects of previous attacks. Similarly, perpetrators can exploit livestreaming functionality when abusing children online. For instance, livestreaming can be used as a way of conducting child sexual abuse by proxy, where children are coerced into abusing themselves or other children in real-time on camera.

- **Recommender systems**: Recommender systems are commonly designed to optimise for user engagement and learn about users' preferences. Where a user is engaging with harmful content such as hate speech or content which promotes suicide, there is a risk that this might result in ever more of this content being served up to them.

We expect services to think about these risk factors when doing their risk assessments (see Volume 3). As we explain in Volume 4, we have designed a number of the measures in our Codes of Practice to target high-risk service types and functionalities.

The functionalities we describe above are not inherently bad and have important benefits. End-to-end encryption plays an important role in safeguarding privacy online. Pseudonymity and anonymity can allow people to express themselves and engage freely online. In particular, anonymity can be important for historically marginalised groups such as members of the LGBTQ+ community who wish to talk openly about their sexuality or explore gender identity without fear of discrimination or harassment. Recommender systems benefit internet users by helping them find content which is interesting and relevant to them. The role of the new online safety regulations is not to restrict or prohibit the use of such functionalities, but rather to get services to put in place safeguards which allow users to enjoy the benefits they bring while managing the risks appropriately.

Online harms and the risk factors which cause them are changing all the time as technology develops and society evolves. The recent emergence of generative AI provides a particularly clear example of this. As well as bringing important benefits, generative AI creates new risks. Image-generation models, for example, can be used in some cases to create CSAM. Studies have also highlighted the use of Generative AI to create 'deepfakes' in support of foreign interference campaigns. They have also been used to generate instructions for how to access unlicensed firearms, and how to make explosive materials.

The constant emergence of new risks makes it important that services conduct regular risk assessments. It also makes robust corporate governance particularly important. Where services have good governance arrangement in place with clear accountability for managing risks, they are more likely to detect and appropriately manage emerging risks. In addition to recommending measures to address specific harms, a key focus for us as we take on our new role will therefore be ensuring that services do robust risk assessments and have appropriate governance arrangements in place. We discuss this further in Volume 3.

What input do we want from stakeholders?

- Do you have any comments on Ofcom's assessment of the causes and impacts of online harms? Do you think we have missed anything important in our analysis? Please provide evidence to support your answer.
- Do you have any views about our interpretation of the links between risk factors and different kinds of illegal harm? Please provide evidence to support your answer.

Volume 3: How should services assess the risk of online harms?

7. Introduction to volume 3

In this volume, we explain our proposals about what governance services should put around managing risk, what services should do to assess the risk of illegal harm, and how they can meet their record keeping and reporting duties.

8. Governance and accountability

What is this chapter about?

Governance and accountability processes are key to a service's ability to properly identify and manage online safety risks.

This chapter sets out our proposed recommendations regarding how services should approach governance and accountability in relation to their illegal content duties under the Act. It covers measures related to governance arrangements; senior accountability and responsibility; internal assurance and compliance functions; and staff policies and practices.

For proportionality reasons, we propose that most measures only relate to large and/or multi-risk services.[2] However, we propose that the requirement for a senior accountable officer applies to all services (U2U and search).

What are we proposing?

We are making the following proposals for all services:

- Name a person accountable to the most senior governance body for compliance with illegal content duties and reporting and complaints duties.

We are making the following proposals for all multi-risk services and all large services[3]:

- Written statements of responsibilities for senior members of staff who make decisions related to the management of online safety risks.
- **Track evidence of new kinds of illegal content on their services, and unusual increases in particular kinds of illegal content**, and report this evidence through the relevant governance channels. U2U services should also track and report equivalent changes in the use of the service for the commission or facilitation of **priority offences**.

[2] For further detail, 10please see our Introduction to Volume 4, where we define 'large' and 'multi-risk' services.

[3] This is with the exception of large vertical search services. This is because 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. We also propose to exclude vertical search from the measure relating to reporting annually to the most senior governance body, for the same reasons.

- A Code of Conduct that sets standards and expectations for employees around protecting users from risks of illegal harm.
- That staff involved in the design and operational management of the service are sufficiently trained in the service's approach to compliance.

We are also making the following proposals for large services:

- The most senior body in relation to the service should carry out and record an annual review of risk management activities in relation to online safety, and how developing governance risks are being monitored and managed.
- Large multi-risk services should have an internal monitoring and assurance function[4] to provide independent assurance that measures taken to mitigate and manage the risk of harm to individuals identified in the risks assessment are effective on an on-going basis, reporting to an overall governance body or audit committee.

Why are we proposing this?

Robust governance processes are an effective way of ensuring good risk management and we therefore expect that widespread adoption of such governance processes will make a material contribution to reducing online harm. Although there is the potential for significant costs in some areas, we consider that good governance is sufficiently important and beneficial to justify these costs. We also consider that the costs of deploying good governance to prevent risks from materialising will often be less significant than the costs services would incur remedying risks that have already materialised. Targeting several of these measures at only large and/or multi-risk services will help ensure we are not imposing undue costs on services that pose a low risk of online harm. Many of the services we are targeting will already have existing governance and accountability arrangements which can accommodate these recommendations.

We are not yet making any recommendations regarding external audit requirements, or regarding linking remuneration and bonuses to online safety outcomes due to limitations in currently available evidence that demonstrates the effectiveness and costs of these proposals.

The proposals for organisations that operate large services are designed to be consistent with the operation of a 'three lines of defence' governance model, and can easily be mapped to the first (management), second (risk management and compliance) and third line of defence (internal audit).

What input do we want from stakeholders?

- Do you agree with our proposals in relation to governance and accountability measures in the illegal content Codes of Practice? Please provide underlying arguments and evidence of efficacy or risks to support your view.
- Do you agree with the types of services that we propose the governance and accountability measures should apply to?
- Are you aware of any additional evidence of the efficacy, costs and risks associated with a potential future measure to requiring services to have measures to mitigate and manage illegal content risks audited by an independent third-party?

[4] Where appropriate, this could be fulfilled by an existing internal audit function.

- Are you aware of any additional evidence of the efficacy, costs and risks associated with a potential future measure to tie remuneration for senior managers to positive online safety outcomes?

9. Services' risk assessment

What is this chapter about?

This chapter covers our guidance about how services can fulfil their duties to assess risks (the 'Risk Assessment Guidance'), including our proposals for the process services should follow when doing their risk assessment and the types of evidence they should consider.

What are we proposing?

We are making the following proposals for all U2U and search services:

- We will guide services to follow a four-step risk assessment process as the best way to ensure that their assessments are 'suitable and sufficient'. These four steps are: (i) understand the harms that need to be assessed; (ii) assess risks by considering the likelihood and potential impact of harms occurring on their service; (iii) implement safety measures and record outcomes of the risk assessment; and (iv) report, review and update the risk assessment.
- We will provide tables listing risk factors, which set out an explanation of what harms these risk factors are associated with and how these increase risks of harm. We call these 'Risk Profiles'. Services should consult these tables when doing their risk assessment. The information in risk profiles is extracted from our assessment of the causes and impact of harms (see above).
- **We will guide all services to consider the following evidence when doing their risk assessment:** Risk Profiles (and relevant parts of Ofcom's Register of Risks), user reports, user complaints, user data including age (where relevant), retrospective analysis of incidents of harm and other relevant information that a service holds.
- Where this evidence does not provide services with a sufficiently good understanding of their risk levels, Ofcom will recommend services look at some or all of the following pieces of additional evidence: results of product testing, results of content moderation systems, consultation with internal experts on risks and safety measures, views of independent experts, internal and external commissioned research, outcomes of external audit or other risk assurance processes, consultation with users and user research, and engagement with relevant representative groups.
- **We will recommend services have a written policy in place to review their assessment at least every 12 months**, and to name a responsible person for overseeing this process (this links to the governance measures in Ofcom's Code of Practice).
- We will recommend that services update their risk assessment whenever a 'significant change' to their service occurs and will provide general principles on how services should interpret what constitutes a significant change. These principles will recognise the importance of the size of a service when considering if a proposed change may be 'significant'.

Why are we proposing this?

This approach reflects our understanding of best practice and current standards in risk management, and mirrors risk assessment processes that have been successfully implemented in other sectors. As explained above, we consider good risk assessment and management will make a material contribution to reducing online harm and that the costs of identifying and managing risks upfront will often be lower than the costs of remedying online harm after the fact. This approach is likely to be complementary to any risk management system that services already have in place, which will reduce the costs of our proposals and ensure they are proportionate.

What input do we want from stakeholders?

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

Specifically, we would also appreciate evidence from **regulated services** on the following:

- Do you think the four-step risk assessment process and the Risk Profiles are useful models to help services navigate and comply with their wider obligations under the Act?
- Are the Risk Profiles sufficiently clear and do you think the information provided on risk factors will help you understand the risks on your service?[5]

10. Record keeping and review guidance

What is this chapter about?

Providers of regulated U2U and search services have duties to make and keep written records of their risk assessments and the measures they take to comply with several duties set out in the Act, as well as regularly reviewing their compliance with relevant duties specified in the Act. This chapter introduces our proposed guidance about how services can fulfil these duties.

What are we proposing?

We are making the following proposals for all U2U and search services:

- Written records can be made and kept in a durable medium of the service's choice.
- Where reasonably practicable, written records should be kept in English (or for services based in Wales, in English or Welsh.
- Written records are written in as simple and clear language as possible.
- A written record must be kept of current risk assessments and compliance measures and must be updated whenever a significant change is made.
- There are additional record-keeping requirements if the service takes alternative measures to those set out in Ofcom's Code of Practice.
- Written records should be retained in accordance with the service's record retention policies, or a minimum of five years, whichever is the longer.

[5] If you have comments or input related the links between different kinds of illegal harm and risk factors, please refer to Volume 2: Chapter 5 Summary of the causes and impacts of online harm).

- Reviews should be scheduled by services and occur with a frequency that allows for a continuous cycle of implementation, monitoring and review.
- **Our expectation is that services should undertake a compliance review at least once a year,** but more frequent reviews may be appropriate if the regulated service becomes aware of compliance concerns or implements new measures. Services should also carry out a compliance review if there is a significant change to any aspect of the design or operation of the service.

We are not proposing to exercise our power to exempt specified descriptions of services from the record keeping and review duty.

Why are we proposing this?

Our proposed guidance seeks to strike a proportionate balance between: accommodating the wide variety of services captured by the guidance; the need for Ofcom to have easy-to-understand, clear and sufficiently detailed written records; and minimising any unnecessary cost or burden on services.

What input do we want from stakeholders?

- Do you have any comments on our draft record keeping and review guidance?
- Do you agree with our proposal not to exercise our power to exempt specified descriptions of services from the record keeping and review duty for the moment?

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

11. Introduction to volume 4: 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.[6]

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.[7]

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

[6] 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.

[7] 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.

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 our 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:

- 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?[8]
- Do you have any comments on the costs assumptions set out in Annex 14, which we used for calculating the costs of various measures?

12. Content moderation (User-to-User)

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

[8] Please see Annexes 7 and 8 to find our draft Codes of Practice.

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.

13. Search moderation (Search)

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.

While 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.

14. Automated content moderation (User-to- User)

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.[9] 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

[9] Though we note there are limits to what they can achieve, in the context of eradicating CSAM online.

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.

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[10] 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.

[10] Fuzzy matching can allow a match between U2U content and a URL list, despite the text not being exactly the same.

15. Automated search moderation (Search)

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.

16. User reporting and complaints (U2U and search)

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.

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)[11], 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.

[11] 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.

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?

18. Default settings and user support for child users (U2U)

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 the first direct communication with that user and of the options available to take 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?

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-platforms 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[12]:

- 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 helping users find content they enjoy and wish to engage with. However, where illegal content is

[12] CSAM; extreme pornography; intimate image abuse; foreign interference; terrorism; encouraging or assisting suicide or serious self-harm; hate; harassment, stalking, threats and abuse.

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?

- 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?

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,[13] have user profiles and have at least one of the functionalities listed at the following footnote:[14]

[13] Coercive and controlling behaviour; harassment, stalking, threats and abuse; hate; grooming; encouraging or assisting suicide or serious self-harm.

[14] User connections; posting content; or user communication (including but not limited to direct messaging and commenting on content).

- 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[15] 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 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?

[15] Harassment, stalking, threats and abuse; hate; grooming; or encouraging or assisting suicide or serious self-harm.

21. User access to services (U2U)

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?

22. Service design and user support (Search)

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.

23. Cumulative Assessment

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?

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?

Volume 5: how to judge whether content is illegal or not? (Illegal Content Judgements Guidance)

25. Introduction to volume 5

In this volume, we set out the approach we have taken to developing our 'illegal content judgements guidance', which explains to services how they should assess whether content is illegal or not.

26. The Illegal Content Judgements Guidance (ICJG)

What is this chapter about?

The Act requires us to provide guidance to services about how they can judge whether a piece of content is likely to be illegal. In this chapter, we set out our proposed high-level approach to developing this Illegal Content Judgements Guidance ('ICJG'). We explain key terms relevant to illegal content judgements and key factors we considered when drafting the ICJG. We then set out the more detailed policy and legal considerations we have had to take into account when developing this guidance for specific offences.

What are we proposing?

The Act requires services to take action against content where they have reasonable grounds to infer that it is illegal. Broadly speaking there are two ways services can meet this duty. If they wish to, they can follow the process set out in our ICJG to determine when there are reasonable grounds to infer that a piece of content is illegal. Alternatively, they can draft their own terms and conditions in such a way that at a minimum all content which would be illegal in the UK is prohibited on their service for UK users and make content moderation decisions based on their terms and conditions.[16] In practice we expect that many services will take the second of these approaches, or a hybrid approach.

In the ICJG we are proposing to provide guidance to services to give them greater clarity about how they should assess whether content is illegal or not. The proposed guidance does not look at whether content may facilitate the commission of an offence. In our proposed guidance, we also set out our provisional view on: (a) what a service should consider to determine if it has 'reasonable grounds to infer that content is illegal content', and (b) what may constitute information that is 'reasonably available' to services when making an illegal content judgement.

[16] Services are free to take down content above and beyond what is illegal under the Act, so long as they make this clear in their terms of service, and that their content moderation practices result in the timely removal of illegal content as set out in the illegal content safety duties.

What input do we want from stakeholders?

- Do you agree with our proposals, including the detail of the drafting? What are the underlying arguments and evidence that inform your view.
- Do you consider the guidance to be sufficiently accessible, particularly for services with limited access to legal expertise?
- What do you think of our assessment of what information is reasonably available and relevant to illegal content judgements?

Volume 6: information gathering and enforcement powers, and approach to supervision

27. Introduction to volume 6

This volume sets out our proposed approach to our information powers, enforcement powers, and our initial thoughts on how we will approach the supervision of certain services.

28. Information powers

What is this chapter about?

The Act gives Ofcom the power to require information we need for purposes of exercising, or deciding whether to exercise, our online safety duties and functions. This chapter gives an overview of Ofcom's information gathering powers and Ofcom's approach to information gathering under the Act.

What are we proposing?

We will use our information gathering powers in a way that is proportionate to the use to which the information will be put and will only issue an information notice where we require information to exercise an online safety function or to decide whether to do so. We expect to use our power to issue statutory information notices regularly from the outset of the regime. Any information notices we issue will clearly set out the purpose of the request and why we require the information. We do not anticipate using our other information gathering powers such as skilled persons reports and powers of entry, inspection and audit as often, and these will typically be reserved for more serious cases.

We have lots of experience in handling information received from regulated services, and other third parties. We will not disclose confidential information unless there is a legal reason to do so, and we will carefully consider the need to disclose against any confidentiality concerns the person providing the information may have.

We expect to publish guidance on how we will use our information gathering powers at a later stage in the implementation of the Act.

Why are we proposing this?

The ability to gather information is fundamental to Ofcom being able to carry out our functions and protect users online. We will therefore use these powers where we think it is proportionate to do so.

What input do we want from stakeholders?

- Do you have any comments on our proposed approach to information gathering powers under the Act?

29. Enforcement powers

What is this chapter about?

This Chapter explains our general approach to regulatory enforcement, how we expect to approach enforcement under the Act and introduces our Online Safety Enforcement Guidance.

What are we proposing?

The Act grants Ofcom a range of enforcement powers and requires us to publish guidance on how we will exercise them. We are consulting on draft Online Safety Enforcement Guidance that sets out how we will normally approach enforcement under the Act. The approach has been informed by our experience and track record of enforcement in other sectors that we regulate.

We may decide to take enforcement action in the interests of citizens and consumers, for example to drive compliance, deter future wrongdoing, protect users from harm and hold wrongdoers to account. If we consider it appropriate to use our statutory enforcement powers under the Act, we will conduct an investigation into the potential breach following the processes set out in the draft Online Safety Enforcement Guidance. This may lead to us issuing a decision on whether a regulatory breach has taken place and imposing financial penalties and other sanctions.

The Act sets out which of the duties on regulated services are subject to enforcement action by Ofcom. As soon as an enforceable duty comes into effect, Ofcom may choose to use the relevant enforcement powers provided in the Act against any service that fails to comply.

Some of the duties, such as the duty to comply with information notices, came into effect at the time the Act passed. Other enforceable duties will not take effect until after Ofcom has finalised the relevant corresponding Codes of Practice or guidance in relation to those duties, or until secondary legislation has been passed.

We expect services to take action to come into compliance with the duties as soon as they take effect. For some duties, we will expect services to comply fully straight away. For example, the duty to comply with information notices, or the duties around risk assessments or child access assessments which already have built in statutory timescales for compliance set out in the Act.

We recognise that when the illegal content and child safety duties come into effect (following Codes of Practice being published), it may take some time for services to put in place all the necessary mitigations. For example, it may be reasonable for services to focus early efforts on putting in place the mitigations that are most likely to protect users from the most serious potential harms, or on mitigations that are relatively quick or simple to implement.

We will take a reasonable and proportionate approach to the exercise of our enforcement powers, in line with our general approach to enforcement and recognising the challenges facing services as they adapt to their new duties. For the illegal content and child safety duties, we would expect to prioritise only serious breaches for enforcement action in the very early stages of the regime, to allow services a reasonable opportunity to come into compliance. For example, this might include where there appears to be a very significant risk of serious and ongoing harm to UK users, and to children in particular. While we will consider what is reasonable on a case-by-case basis, all services should expect to be held to full compliance within six months of the relevant safety duty coming into effect.

Why are we proposing this?

The Act grants Ofcom a range of enforcement powers and requires us to publish guidance on how we will exercise them.

What input do we want from stakeholders?

- Do you have any comments on our draft Online Safety Enforcement Guidance?

30. Supervision

What is this chapter about?

This chapter sets out our approach to supervision of a small subset of the highest reach or highest risk services in scope of the Online Safety Act. Supervision will help ensure that these services have appropriate systems and processes to achieve the key outcomes intended by the Act to make life safer online for people across the UK.